



Informatics

Choice-Based Preference Elicitation to Reduce the Cold Start Problem of a Leisure Activities Recommender in a Mobile App

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Medieninformatik

eingereicht von

Andreas Fink, Bakk.

Matrikelnummer 00001404

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assistant Prof. Mag.a rer.nat. Dr.in techn. Julia Neidhardt

Wien, 2. Mai 2025

Andreas Fink

Julia Neidhardt



Informatics

Choice-Based Preference Elicitation to Reduce the Cold Start Problem of a Leisure Activities Recommender in a Mobile App

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Andreas Fink, Bakk.

Registration Number 00001404

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Mag.a rer.nat. Dr.in techn. Julia Neidhardt

Vienna, May 2, 2025

Andreas Fink

Julia Neidhardt

Erklärung zur Verfassung der Arbeit

Andreas Fink, Bakk.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, habe ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT-Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 2. Mai 2025

Andreas Fink

Danksagung

Ich möchte meiner Betreuerin, Assistant Prof. Mag.a rer.nat. Dr.in techn. Julia Neidhardt, meinen Dank aussprechen. Trotz der doch etwas längeren Umsetzungsphase hat sie mir stets wertvollen fachlichen Input gegeben und bis zum Abschluss das Interesse an diesem Forschungsprojekt nicht verloren. Ein besonderer Dank gilt meinen Eltern, die mir überhaupt erst ermöglicht haben, ein Studium aufzunehmen. Ebenso danke ich allen, die mich in irgendeiner Weise bei der Durchführung unterstützt haben – sei es auch "nur" durch aufmunternde Worte, die sich in entscheidenden Momenten als ebenso wichtig erwiesen haben. Besonders hervorheben möchte ich Fabian, der sich insbesondere in den letzten Monaten des Schreibprozesses – mal mehr, mal weniger erfolgreich – als Motivator ausgezeichnet hat. Schließlich danke ich Marlene, die über die Jahre hinweg mehrere Versuche, mein Studium abzuschließen, miterlebt hat und mich dabei dennoch immer unterstützt hat.

Acknowledgements

I would like to express my gratitude to my supervisor, Assistant Prof. Mag.a rer.nat. Dr.in techn. Julia Neidhardt, who, despite the rather extended implementation period, consistently provided valuable academic input and never lost interest in the research project until its completion. Special thanks are due to my parents, without whose support I would not have had the opportunity to pursue academic studies in the first place. I also extend my thanks to everyone who supported me in any way during the course of this project – even if it was "only" through words of encouragement, which at times proved just as essential. Particular appreciation goes to Fabian, who, especially in the final months of writing, distinguished himself – sometimes more successfully, sometimes less so – as a source of motivation. Finally, I would like to thank Marlene, who has witnessed several of my attempts to complete this degree over the years, yet continued her unwavering support throughout.

Kurzfassung

Frühe Phasen von Nutzerinteraktionen stellen für Empfehlungssysteme eine große Herausforderung dar, insbesondere aufgrund des Kaltstartproblems. Dieses Problem entsteht dadurch, dass neue Nutzer oder Angebote ohne historische Daten keine personalisierten Empfehlungen erhalten können. Diese Arbeit untersucht die Effektivität von vier visuellen Methoden – Swipe, Rating, Two-Items und Four-Items – zur Ermittlung von Interessen, um ein initiales Nutzerprofil während des Onboardings in einem Empfehlungssystem für Freizeitaktivitäten zu erstellen. Dieses Empfehlungssystem ist in einer mobilen App integriert.

Dazu wurde ein browserbasierter Prototyp einer Umfrage entwickelt und die Methoden hinsichtlich Abschlussrate, Zeiteffizienz, Benutzerfreundlichkeit und Profilgenauigkeit bewertet. Die Studie umfasste 382 Teilnehmer, welche die Umfrage vollständig durchgeführt haben.

Die Ergebnisse zeigen klare Kompromisse zwischen der Einfachheit der Interaktion und der daraus resultierenden Qualität des Präferenzprofils. Die Swipe-Methode erzielte die höchsten Werte bei Benutzerfreundlichkeit und Abschlussrate, führte jedoch zu den ungenauesten Präferenzprofilen. Die Rating-Methode liefert die genauesten Präferenzprofile, jedoch dauerte die Durchführung länger und die Abschlussrate war geringer als bei der Swipe-Methode. Die Two-Items-Methode zeigte eine ausgewogene Leistung hinsichtlich Zeiteffizienz und Profilgenauigkeit, hatte jedoch die niedrigste Abschlussrate. Die Four-Items-Methode konnte zwar bei der Abschlussrate überzeugen, die Profilgenauigkeit lag aber unter der Rating-Methode und der Two-Items-Methode. Zudem war die Durchführungszeit der Four-Items-Methode signifikant länger im Vergleich zu den anderen Methoden.

Diese Arbeit leistet einen Beitrag zur Reduzierung des Kaltstartproblems in Empfehlungssystemen und bietet wertvolle Einblicke in die Gestaltung benutzerfreundlicher und effektiver Präferenzabfragemethoden im Kontext einer mobilen App für Freizeitaktivitäten.

Abstract

Early stages of user interactions pose a significant challenge for recommender systems, particularly due to the cold start problem. This issue arises when new users or items lack historical data, making it difficult to generate personalized recommendations. This thesis investigates the effectiveness of four visual preference elicitation methods – Swipe, Rating, Two-Items, and Four-Items – for constructing an initial preference profile during the onboarding process of a mobile leisure activities recommender system.

For this purpose, a browser-based survey prototype was developed, and these methods were evaluated in terms of completion rate, time efficiency, usability, and profile accuracy. The study included 382 participants who completed the survey.

The results show clear trade-offs between ease of use and the quality of the resulting preference profile. The Swipe method achieved the highest scores in usability and completion rate, but resulted in the least accurate preference profiles. The Rating method produced the most accurate profiles, but required more time and showed a lower completion rate than the Swipe method. The Two-Items method showed a balanced performance in terms of time efficiency and profile accuracy, but had the lowest completion rate. While the Four-Items method performed well in terms of completion, its profile accuracy was lower than that of the Rating method and the Two-Items method. Additionally, the completion time was significantly longer compared to the other methods.

This thesis contributes to the mitigation of the cold start problem in recommender system and provides valuable insights into the design of user-friendly and effective preference elicitation methods in the context of a mobile leisure activities application.

Contents

Kurzfassung	xi
Abstract	xiii
Contents	xv
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement	2
1.3 Research Questions and Research Objectives	2
1.4 Methodological Approach	4
1.5 Structure of the Thesis	6
2 State of the Art	7
2.1 Recommender Systems	7
2.2 Cold Start Problem	9
2.3 Preference Elicitation as a Solution to CSP	10
2.4 Visual Preference Elicitation in Tourism Recommender Systems	11
3 Methodology	13
3.1 Visual Elicitation Methods	13
3.2 Survey Design and Questionnaire	15
3.3 Image Selection and Item Pool	17
3.4 Preference Profile Calculation	18
3.5 Evaluation and Participant Dimensions	21
3.6 Data Collection	23
4 Design and Implementation	25
4.1 Artefact Architecture and System Components	25
4.2 User Interface Design	26
4.3 Implementation of Elicitation Methods	27
4.4 Item Selection and Control Logic	28
4.5 Iterative Refinement of the Prototype	31
	xv

5	Evaluation	33
5.1	Participants	33
5.2	Completion Rate	35
5.3	Time Efficiency	35
5.4	Usability	40
5.5	Profile Fit	43
5.6	Effectiveness of Swipe-Based Onboarding on User Retention and Conversion	47
6	Discussion	51
6.1	Comparative Analysis of Elicitation Methods	51
6.2	Comparison With Related Work	52
6.3	Limitations	53
7	Conclusion	55
7.1	Summary	55
7.2	Contributions	56
7.3	Future Work	57
A	Figures	59
	Overview of Generative AI Tools Used	69
	List of Figures	71
	List of Tables	73
	Glossary	75
	Acronyms	77
	Bibliography	79



Introduction

1.1 Motivation

Recommender system (RS) are widely used in the digital environment, providing personalized content [GW15], advertisements [TV11], and product suggestions [YLS⁺11]. These systems analyze user behavior to create profiles [AT05] and offer recommendations [Bur00]. In this context, companies aim to enhance user satisfaction, engagement, and platform usage, ultimately increasing revenue [TV11].

Recommendations typically rely on similarities between users or items and fall into three categories: content-based filtering (CBF), collaborative filtering (CF) and hybrid approaches. CBF analyze user interactions to create profiles, while CF identifies similarities in user preferences, often using statistical calculations based on ratings or click patterns. Hybrid approaches combine CBF and CF techniques.

Despite the multiple benefits of RS, they face multiple challenges [TV11]. One of the main concerns is to provide suitable suggestions for new offers and new users for which no historical data is yet available, which is known as the cold start problem (CSP).

Especially when recommending leisure activities, the CSP proves to be a particular challenge. Unlike other domains such as music or movie recommendations, leisure activities are often associated with higher costs and a greater time investment [SNW20]. It is therefore crucial to offer accurate and personalized recommendations from the start in order to increase user satisfaction and foster long-term engagement.

Recent studies indicate that people attach an ever-increasing importance to work-life balance. In particular, the significance attributed to leisure time has soared after the limitation of the global COVID-19 pandemic [CPB21]. The perceived importance of recreational activities makes the development of a RS for this field a challenging task [SNW20].

1.2 Problem Statement

The CSP is a major challenge in the development of a RS for an application for recommending leisure activities. The application aims to quickly and effectively engage new users and provide them with personalized recommendations despite the lack of sufficient user history. The problem is further complicated by the variety and ever-changing range of leisure activities offered in the application and the subjective and personalized nature of users' preferences and needs.

Surveys are a popular approach to quickly get information from users about their preferences to reduce the CSP. A survey can be displayed to users when they first visit a platform or during use in between. A certain number of items are suggested, and users rate them according to their preferences. A distinction can be made between several types of ratings. For example:

- unary (I like it)
- binary (I like it / I don't like it)
- ordinal (e.g., most preferred, preferred, neither preferred nor not preferred, not preferred, not at all preferred)

Another method is choice-based elicitation. Here, users are shown e.g., two items and they indicate which one they prefer [JBB11].

The problem is to identify users' preferences and needs without inconveniencing them or confronting them with boring questionnaires. Especially in the context of a mobile app, it is important to engage users in a fun and engaging way to get accurate information and create a positive onboarding experience. The selection of items for the survey plays a crucial role in this, as visual stimuli can play an important role in decision-making and preference formation [NSSW15]. To bind potential users, the selection process needs to be user-friendly and should encourage users to complete the full process. Previous work shows that in the digital environment the attention span of humans is only 8 seconds anymore [SS17].

1.3 Research Questions and Research Objectives

This study investigates how different visual preference elicitation methods perform in a simulated onboarding setting of a mobile leisure RS. Specifically, it addresses the following research questions:

Research Question 1 (RQ1): Which visual preference elicitation method achieves the highest performance in terms of completion rate, time efficiency, usability, and profile fit during onboarding in a mobile recommender system for leisure activities?

Research Question 2 (RQ2): To what extent does the integration of a visual preference elicitation method during onboarding affect user retention and conversion rates over a two-month period in a mobile leisure activities app, compared to a baseline without visual onboarding?

- **Retention Rate:** Defined as the percentage of users who continue using the app over time.
- **Conversion Rate:** Quantified as the percentage of users who engage in key activities, such as visiting detail pages or making purchases.

To address these research questions, four preference elicitation methods were selected based on their prior application in related research and their suitability for mobile application contexts:

- The first method uses a binary rating where users can indicate whether they like or dislike an activity by clicking on two defined icons or by gesture control.
- The second method is based on a star-based rating, where users have a scale from 1 to 5 to express their preferences for different activities.
- The third method is based on a choice-based elicitation with two items. Users are presented two activities and have to choose which one they like better.
- The fourth method also uses a choice-based elicitation, but with four items. Users are presented four activities and have to choose which they like best and which they like least.

Each method uses visual representations, such as pictures, to display the items for the survey. For this purpose, taxonomies must be developed on the basis of which the images are displayed in the respective step.

In the context of a research project supported by the Österreichische Forschungsförderungsgesellschaft mbH (FFG), a leisure activities app from a corporate partner is used as a use case. In this app, the activities on offer are divided into six overarching categories: *Sports*, *Entertainment*, *Cultural activities*, *Animal-related activities*, *Course* and *Relaxation*.

To classify participant feedback effectively, a selection of corresponding images must be made for visualization in each of the predefined categories.

By answering the outlined research questions, this thesis aims to provide valuable insights into the effectiveness of visual preference elicitation methods in mitigating the CSP within a mobile leisure RS. The investigated methods are designed to capture user preferences in a non-intrusive and engaging way, combining visual stimuli with intuitive interaction mechanisms to enhance the onboarding experience. The findings contribute to a better understanding of how different elicitation strategies influence user satisfaction,

recommendation accuracy, and long-term engagement. Furthermore, the development and empirical evaluation of a browser-based prototype serve as an additional contribution to both practice and scientific research in the field of human-centered recommender systems.

1.4 Methodological Approach

In this thesis, the Design Science Research (DSR) model by Hevner [Hev07] serves as the methodological foundation, which structures the research process into a relevance cycle, a rigor cycle, and a design cycle. This methodological framework supports the development of an artifact that responds to a real-world challenge while also contributing to academic discourse. In this study, the artifact consists of a set of visual preference elicitation methods aimed at improving the onboarding in a mobile RS for leisure activities by reducing the cold start problem. The three DSR cycles are described in the following subsections:

1.4.1 Relevance Cycle

In the relevance cycle, research is linked to the practical application context. The cold start problem in the domain of leisure activity recommendations serves as the core issue. The problem context was explored through a structured collaboration with the corporate partner in the leisure sector, whose mobile application provides practical relevance to this study. A key requirement in this context is to develop onboarding methods that not only capture user preferences, but are also intuitive, visually appealing, and optimized for mobile use.

The relevance cycle concludes with the integration of these methods into an interactive survey-based prototype, which serves as the testing ground for empirical evaluation.

1.4.2 Rigor Cycle

The rigor cycle connects the research to the state of the art and uses established theories, models, and prior artifacts from related domains such as RSs, preference elicitation, and human-computer interaction. A literature review was conducted, covering areas such as cold start strategies in RS, visual approaches to preference elicitation, as well as established methods in choice-based decision modeling. Furthermore, existing taxonomies of leisure activities and psychological models of user preference [NSSW15] were considered in the construction of the image-based elicitation content. Each method was designed based on prior empirical insights into usability, decision effort, and preference stability.

The knowledge gained from this review guided the design of the methods and provided the scientific foundation needed to ensure that the developed solutions contribute beyond routine IT implementation. The rigor cycle also ensures that the outcomes of the research are properly documented and returned to the academic community through this thesis.

1.4.3 Design Cycle

The main focus is on the design cycle, which integrates findings from the relevance and rigor cycle in order to iteratively build and evaluate the artefact. In this thesis, the design cycle consists of several key steps:

Prototyping the Elicitation Methods

Each of the four methods was implemented using a shared visual design language, with leisure activity items represented through images and categorized according to a taxonomy of six first-order leisure domains: Sports, Entertainment, Cultural activities, Animal-related activities, Course and Relaxation.

Study Design and Path Selection

To reduce cognitive load while ensuring method comparability, participants were assigned to one of three paths, each comparing a pair of methods. The goal was to elicit enough preference data while maintaining engagement throughout the onboarding process.

Data Collection and Profile Generation

Users interacted with the selected methods, and the resulting preference profiles were automatically calculated. Participants were then asked to evaluate the relevance and accuracy of these profiles.

Quantitative Evaluation

A quantitative evaluation was conducted to measure the performance and effectiveness of the implemented preference elicitation methods. The evaluation focused on multiple dimensions. The completion rate served as an indicator of overall participant engagement, reflecting the extent to which users were willing and able to complete the onboarding process using the respective methods. Time efficiency was measured by recording the duration required to complete each method, providing insights into their practical feasibility and cognitive demand in a mobile context. In addition, usability was evaluated through post-task questionnaires, capturing participants' subjective impressions regarding choice difficulty and ease of use with the interaction process. The resulting user profiles were evaluated based on perceived accuracy, with participants assessing how well the elicited preferences aligned with their actual interests. To capture longer-term effects, retention and conversion rates were monitored over a two-month period after implementation. Retention measures the percentage of users who keep using the app after the initial onboarding phase, serving as an indicator of sustained user engagement and satisfaction. Conversion, in contrast, tracks the percentage of users who performed desired actions within the app, such as visiting activity detail pages or making a booking, showing how well the onboarding process encourages users to deeper interaction.

Through repeated cycles of testing and refinement, the artifact was improved based on user feedback.

1.5 Structure of the Thesis

This thesis is structured in seven chapters, which step by step lead from the problem definition through the methodological implementation to the evaluation and discussion of the results.

The thesis begins by introducing the topic, describing the motivation, the research problem, the research questions, and providing an overview of the methodological approach.

Building upon this introduction, the subsequent chapter presents the state of the art. It discusses relevant fundamentals on RSs, the CSP, and various preference elicitation approaches. A particular focus is placed on choice-based and visual methods in the field of tourism applications.

The methodological approach is then described, including the research design, including the development of the prototype, and the design of the preference elicitation methods. It also covers the procedure for data collection, and the definition of evaluation criteria.

This is followed by Chapter 4 focusing on the implementation of the prototype. It outlines the technical and visual realization of the prototype, the implementation of the four elicitation methods, and the iterative development process.

Next, the results of the evaluation are presented. The chapter introduces the sample and assesses the defined metrics, including completion rate, time efficiency, usability, and profile fit. Furthermore, the influence of a preference elicitation method on retention rate and conversion rate is examined.

Chapter 6 discusses the findings in light of the existing literature, and relating them to the research questions. It draws theoretical and practical implications and reflects on limitations of the study as well as on lessons learned.

The thesis concludes by summarizing the key findings and offering an outlook on future research directions.

CHAPTER 2

State of the Art

This chapter provides an overview of the current state of research in the field of RSs, with a particular focus on strategies addressing the CSP. The chapter is structured into four main sections and begins by introducing fundamental concepts and mechanisms of RSs. The following section discusses the challenges caused by the CSP and explores algorithmic solutions for its mitigation. Section 3 highlights the role of explicit preference elicitation methods. Finally, visual techniques are investigated, particularly in tourism and leisure applications, as an engaging approach to preference collection.

2.1 Recommender Systems

Over the past three decades, the Internet has become a central medium for information retrieval, communication, and e-commerce [Agg16]. With the continuous increasing number of offers, users are confronted with choice overload due to the unmanageable amount of content. This overload makes it difficult for users to find relevant and useful content [RD22]. In order to address this data overload, solutions have been developed, including RSs. RSs are software tools and techniques that suggest personalized content recommendations to users in form of a ranked list [JZ12]. They aim to estimate the probability that given content is relevant to users based on their interests. In order to make these predictions, users' historical behavior and expressed interests are analyzed. This can be achieved through implicit and explicit feedback. Implicit feedback involves passively collected data that reflects user behavior, e.g. viewing the detail page of a product, adding items to a wishlist, or completing a purchase. In contrast, explicit feedback is gathered directly from users, for example through rating systems, surveys, or preference questionnaires, where users actively indicate their likes and dislikes. Based on this collected feedback, a preference profile of the respective user is calculated [RRS10]. This preference profile is the representation of a user's interests, behavior and characteristics, e.g., age, gender, or location [MID23] and is one of the core components of a RS to provide

personalized content recommendations. Beyond enabling personalization, a preference profile allows more accurate and relevant, as well as more efficient recommendations, as content that does not align with the user's interests can be effectively filtered out. A key aspect of the effectiveness of a preference profile is its profile accuracy. Profile accuracy refers to the degree to which the calculated profile correctly reflects the true preferences and interests of the user. High profile accuracy is critical, as it directly influences the quality and acceptance of the recommended content. In contrast, low profile accuracy can lead to irrelevant suggestions, resulting in user dissatisfaction and decreased engagement. Since user preferences might change over time, the preference profile should be recalculated regularly to maintain a high level of accuracy.

There are several methods with different strengths and use cases for calculating the preference profile. In context of the development and implementation of RSs, they can be distinguished into three broad categories: CF, CBF and hybrid approaches.

First, CF is founded on the idea that users who have selected certain items in the past are likely to do so again in the future. Based on this assumption, a distinction can be made between user-based and item-based CF. User-based CF recommends items to a user based on the preferences of other users with similar tastes. In contrast, item-based CF analyzes a user's past interactions to suggest items that are similar to those the user has previously engaged with [RRS10]. Both types faces several challenges, such as the sparsity problem, where the lack of a sufficient number of ratings makes it difficult to make accurate recommendations. They are also vulnerable to the CSP, which occurs when there is not enough historical data for new users or items.

Second, CBF recommends items based on the attributes of items a user has previously interacted with. For instance, a user who regularly watches documentaries might be recommended other documentaries, even though there are no other users with similar behavior. However, this approach requires detailed metadata about item characteristics and bears the risk of over-specialisation. In such cases, users may only be presented with content that is similar to their past choices, limiting opportunities for serendipitous discovery of new and potentially interesting items.

Finally, hybrid RSs combine collaborative and content-based filtering to leverage the strengths of both approaches while minimising their weaknesses. For example, a hybrid RS could use content-based filtering to suggest items for new users and switch to collaborative filtering when sufficient historical data is available to make recommendations based on user preferences [Bur07].

Regardless of the underlying method, the quality of the recommendations remains a key role in the success of a RS. Not only do they improve the user experience by suggesting content precisely tailored to individual preferences, but they also enhance user satisfaction. Personalized recommendations engage users by consistently presenting content that aligns with their interests. In addition, RSs facilitate the navigation through the large amount of available content by highlighting the items that match the user's preferences. This leads to higher conversion rates, as users are more willing to consume the suggested

content, which in turn increases the revenue of the respective platform operator.

2.2 Cold Start Problem

While RSs offer significant benefits, they often struggle with limited data availability in the early lifecycles of users or items. This challenge is known as the CSP and can further be categorized in user cold start (UCS), item cold start (ICS) and system cold start (SCS). This section explores the types of this problem, its consequences, and possible algorithmic solutions.

The term UCS is used when new users have had few or no interactions with a given system. A similar problem is the ICS, where new items are added to a system without any previous user interaction. SCS occurs when a RS itself is newly introduced and has neither user nor item historical data. These scenarios limit the ability to build accurate user profiles, leading to lower recommendation quality. As a result, several techniques have been proposed to mitigate this issue. Beyond the already presented CBF and hybrid approaches, advanced models such as matrix factorization (MF), community detection, and graph-based techniques have demonstrated considerable potential [GW15] [GV17] [GJ17].

MF techniques, such as Singular Value Decomposition (SVD), are among the most prominent approaches. Users and items are mapped into a shared latent space in order to uncover hidden patterns in interaction data. By considering metadata, e.g. categorical tags, contextual features, or user attributes, into matrix factorization models, this technique allows the system to estimate preferences even with sparse rating data, thereby improving recommendation quality in cold-start scenarios [GW15].

Community detection methods focus on identifying clusters of users who share similar preferences or behaviors. By grouping users into communities, these models infer preferences for new users based on the collective behavior of their community. This reduces dependency on individual user histories and helps mitigate the sparsity problem [GV17].

Graph-based techniques model users and items as nodes in a network, with interactions represented as edges. [GJ17]. By applying algorithms such as label propagation and random walks, latent connections can also be uncovered in sparse datasets [ZYZ⁺22]. This approach makes it possible to close gaps in data and recommend relevant items in early phases of user interaction [WL22].

Recent years have seen a growing interest in deep learning-based approaches. Convolutional neural networks (CNNs) and Recurrent neural networks (RNNs), for instance, can extract latent features from unstructured data such as images or text [MGW21]. CNNs are well-suited for visual content analysis [Wu17], making them useful in domains like tourism, where image-based recommendation is applied. RNNs, on the other hand, are particularly effective in processing sequential data such as clickstreams or text reviews, capturing temporal dependencies and user intent [BB23].

Embedding techniques inspired by models like Word2Vec transform users and items into dense vector spaces where semantic relationships are preserved. These embeddings

enable similarity calculations and clustering, providing a foundation for more nuanced recommendations, even with limited data [MSdGL16].

Transfer learning further enhances RSs by reusing knowledge from related domains or tasks. Pre-trained models are fine-tuned with minimal domain-specific data, enabling the system to generate accurate recommendations based on the knowledge of the original tasks [JGJ⁺22].

Hybrid neural networks combine the strengths of MF and deep learning by integrating latent factor models with deep architectures. Neural collaborative filtering (NCF) is one such approach that uses multilayer perceptrons to model complex user-item interactions. These models are capable of capturing nonlinear relationships and outperform traditional CF models in CSP scenarios [MGW21].

Context-aware systems incorporate situational variables such as time, location, or user mood into the recommendation process [GT11]. For instance, a user may prefer different items in the morning than in the evening, or when traveling versus at home [BA09]. Integrating contextual data improves the relevance and timing of recommendations.

Finally, reinforcement learning and meta-learning strategies offer dynamic and adaptive approaches to recommendation. Reinforcement learning treats the recommendation process as a sequential decision-making task, where the system learns from user feedback to optimize long-term engagement [AG21]. Meta-learning, or "learning to learn," enables rapid adaptation to new users or items by generalizing from prior learning episodes, which makes it particularly well suited for few-shot learning contexts [WYKN20] [LHZZ24].

While these algorithmic approaches are effective, they typically depend on implicit user interaction. Therefore, the following section examines explicit preference elicitation as a more direct and intentional method of collecting user input.

2.3 Preference Elicitation as a Solution to CSP

Explicit preference elicitation provides a user-focused solution to the limitations of implicit data collection, particularly in cold-start scenarios. By collecting direct user feedback through structured interfaces, this approach enhances the quality and personalization of initial recommendations.

One of the most established techniques in preference elicitation is rating-based input, where users assign numerical or categorical scores, such as giving stars on a scale from one to five [JBB11]. Although the simplicity of this approach has been emphasized in prior research, it has also been criticized for potential inconsistencies and cognitive biases, as users often struggle to assign absolute values to their preferences [JBB11] [YB22].

As an alternative, pairwise comparison methods ask users to select their preferred option from two presented items [KPG22]. This approach aligns more closely with natural decision-making processes by shifting the focus from absolute to relative evaluations, thereby reducing cognitive load [GS10]. Building on these principles, choice-based

elicitation methods have been developed, assuming that users generate more consistent and stable preference statements when making relative judgments [JBB11] [CFY⁺21] [JN11]. Louviere et al. [LHS10] demonstrated in their work on Stated Choice Methods that selecting between alternatives requires less cognitive effort than assigning isolated ratings. This contributes to the formation of more reliable and robust preference profiles and explains the increasing use of such methods in the design of personalized RSs. Building on the idea of simplicity and enhanced user engagement, choice-based preference elicitation approaches have proven particularly promising. Loepp et al. [LHZ14] introduced a method where users iteratively choose from a set of items, gradually refining their preferences. These approaches integrate well with latent factor models and studies have shown that choice-based methods are less demanding and deliver more satisfying results compared to traditional rating-based methods [LHZ14] [GW15].

Other interactive formats, such as critique-based elicitation, offer users the ability to iteratively refine their preferences by evaluating sample items through predefined criteria (e.g., "less expensive") [PC08]. Pommeranz et al. [PBW⁺12] emphasize the value of this approach, as this iterative refinement reveals nuanced preferences and allows users to actively influence the recommendation process [CP09].

Taking this concept a step further, Conversational Recommender Systems (CRSs) simulate dialogue-based interactions, dynamically adjusting queries based on user responses [PC08]. With the integration of active learning strategies, CRSs can prioritize the most informative questions, optimizing both user engagement and profile accuracy [ILN⁺21].

Across all these approaches, the design of the user interface plays a critical role, as it must balance cognitive load, user effort, and the level of expressiveness necessary to articulate preferences. The research by Pommeranz et al. [PBW⁺12] highlights the importance of interface design that allows users to fully explore and express their preferences without unnecessary effort. Users are generally more willing to invest time in feedback mechanisms if they feel that the effort directly improves the quality of the recommendations [KWH10] [Kim21].

The next section extends these ideas into the visual domain, illustrating how image-based elicitation methods can be particularly effective in experience-driven environments such as tourism.

2.4 Visual Preference Elicitation in Tourism Recommender Systems

Extending the previously discussed elicitation strategies, visual methods have proven particularly valuable in domains where user decisions are influenced by visual stimuli. Tourism is a prominent example, where visual content strongly influences both expectations and emotions.

Neidhardt and colleagues [NSSW15] introduced a seven-factor model supported by a curated image set for preference elicitation in tourism contexts. Experts assigned up to

seven images to thousands of tourist points of interest (POI), with each image ranked according to its relevance. Based on users' image selections, a profile was generated to reflect their travel preferences. Based on this, Sertkan et al. [SNW20] implemented a web-based profiler that analyzed user-uploaded images using CNN classifiers. These classifiers mapped each image to one or more of the seven factors, allowing the system to compute an aggregated profile vector. Users could then fine-tune their preferences before receiving destination recommendations.

These studies demonstrate that visual elicitation enhances user engagement while also facilitating intuitive, emotion-driven decision-making. By making use of images, systems can bypass the limitations of textual input and better align with users' natural evaluation processes, which is particularly important in leisure and lifestyle contexts.

In conclusion, visual preference elicitation represents a promising direction for addressing cold-start challenges, particularly in domains where emotional factors strongly influence user decision-making.

CHAPTER 3

Methodology

The visual preference elicitation methods developed in this research represent the core artifact of the study. This chapter outlines the methodological steps taken to design, implement, and evaluate these methods, following the principles of the DSR methodology. To contribute to academic research while generating valuable insights for managerial practice, the study conducted in a corporate partner operating in the leisure sector. Specifically, the company's mobile application served as research context. The app was originally developed to support individuals in planning their leisure time activities. After indicating time, date, and location, users are presented with a list of leisure time activities, which could range from cultural events like exhibitions to sporty activities, such as dance classes. The app can be downloaded via the given mobile app stores and is freely accessible to users. At the point of the study, minimal personal information was required to use the app.

To generate a list of activities that represent a good fit between the app's recommendation of leisure time activities and the users' personal preferences, the app relies on a tag-based RS that constructs user profiles based on implicit feedback. Consequently, the CSP outlined in Chapter 2.2 represents a significant challenge for the given application. By implementing an image-based elicitation system, the corporate partner had hoped to improve user experiences. Particularly, in the course of the onboarding process, i.e., when users start using the app for the first time and are therein informed about its features, users can indicate their leisure time preferences to receive more fitting recommendations.

3.1 Visual Elicitation Methods

Given the scope of this thesis, the focus was set on four visual query methods as a baseline for the comparative analysis, which will be described in the following. The selection of these specific methods was based on a variety of criteria, most notably how commonly the methods are applied or have been explored in prior research [JBB11] [GW15] [KM17] [KPG22].

Additionally, the intuitiveness of each method was considered, along with its suitability for mobile implementation.

The Swipe method represented the binary assessment mode where participants indicated whether they liked or did not like an item. The second method used an ordinal scale with values from 1 to 5 (hereafter referred to as Rating method), allowing users to express their preferences for different types of activities by attributing one to five stars. The more stars were given, the higher the preference for the respective option. Whereas the first two methods focused on one specific activity or type of activity, methods three and four followed a comparative approach. Specifically, method three (hereafter referred to as Two-Items method) represented a choice-based task with two items. Users were presented with two activity options and were asked to indicate which of the two they prefer. Finally, in method four (Four-Items method), users were presented with four divergent items and were asked to choose which of the given activities or types of activities they like the most and which one they like the least.

To ensure a comprehensive comparison among query methods, presenting users with all four types of query methods would have been beneficial as study participants would have been asked to assess the suitability, usability, etc. of all types of methods. For a complete comparison, each method would have had to be compared with all the other types. As a result, the survey would have had to have six different paths.

Since the number of participants completing the survey could not be predicted in advance, the decision was made to implement only three distinct paths. This approach aimed to ensure that the number of complete responses per path remained sufficient for detecting a statistically significant difference in the effectiveness of the respective method. In particular, the following paths were selected and implemented:

- Comparison of Swipe method and Four-Items method
- Comparison of Swipe method and Rating method
- Comparison of Four-Items method and Two-Items method

These paths were selected based on the assumption that the data obtained from participants' responses could be used to infer potential outcomes for the paths that were not explicitly included in the study. The Swipe method and Rating method share the characteristic that only a single item is presented for evaluation in each step. In contrast, the Two-Items method and Four-Items method involve displaying multiple items simultaneously, from which participants must make a selection. Consequently, the study design enables a comparative analysis of methods that involve a single item per step, both single and multiple items per step, and exclusively multiple items per step.

The selection of the Swipe method and Four-Items method was guided by considerations of usability and decision clarity. A review of the literature indicated that the Four-Items method, leveraging the Maximum Difference Scaling (MaxDiff) algorithm, is theoretically

expected to yield accurate results [KM17]. However, it also expects to demand higher cognitive effort from participants when making decisions [HSvEB21]. In contrast, binary ratings (i.e. Swipe method) have been previously shown to produce less precise results, but are perceived as more intuitive and easier to use, due to its binary characteristic [SLL⁺13]. Based on these insights, it was hypothesized that the Swipe method and Four-Items method represent two extremes in terms of usability and decision clarity, while the Rating method and Two-Items method fall in between. This design is expected to support the identification of trends and probabilistic inferences regarding the paths not explicitly tested.

3.2 Survey Design and Questionnaire

Due to the high complexity and costs associated with integrating the survey into the mobile app, a browser-based prototype was developed instead. This approach facilitated broader participation by allowing access to the survey through a simple link, thereby removing potential barriers like mandatory app installation.

The survey structure consisted of seven steps, including two preference elicitation methods, corresponding questionnaires for each method, a comparative profile selection, and a final section for optional comments (Figure 3.1). Participants were randomly assigned to one of the three paths based on the least-used condition.

Step 1 A brief introduction was displayed to the participants, providing an overview of the survey's purpose and the estimated completion time. By clicking the button "Start Survey", a new session was initiated. To ensure balanced distribution among the different paths, the system selected the path from the database table *xpo-elicitation-mode* that had been completed the least frequently. The counter for the selected path was incremented by one, and a new entry for the participant was created in the database table *xpo-elicitation-user*. This data record included a reference to the assigned path, ensuring proper tracking of participant assignments.

Step 2 The defined elicitation method for the selected path and assigned position was presented to the participants. Initially, a pop-up window appeared, providing a brief textual explanation of how to complete the elicitation method. This pop-up could be closed by clicking a "X" symbol in the top-right corner or by clicking on the surrounding overlay that dimmed the background. Once the instructions had been dismissed, the first item – or set of items, depending on the method – was immediately displayed for evaluation. During the elicitation process, the participants were shown their progress through a visual step counter (e.g., "6/18") within the currently evaluated method. Additionally, an "i" symbol was displayed in the top-right corner, allowing them to reopen the instructions at any time.

Step 3 The questionnaire was then displayed, consisting of five propositions that allowed participants to evaluate their decision-making process during the substeps of

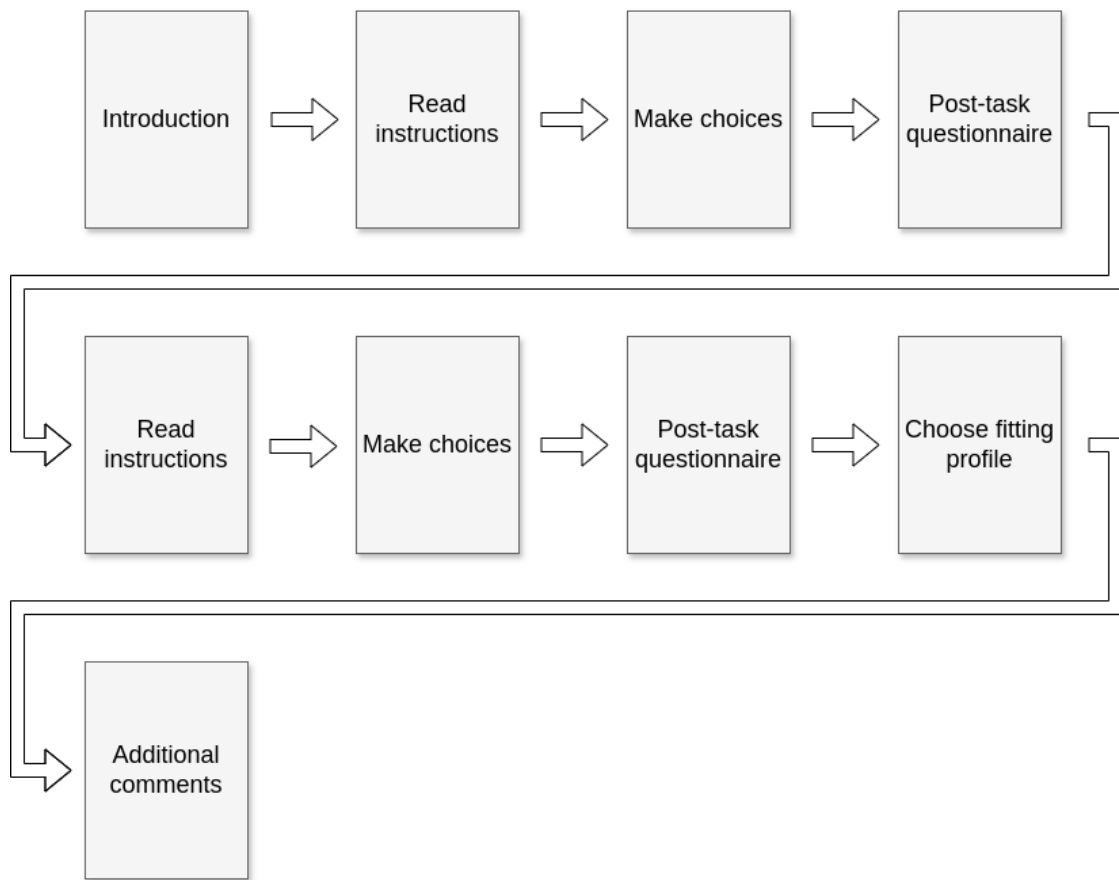


Figure 3.1: Visual representation of the survey flow.

the preceding elicitation method, as well as their overall experience with the method (Figure A.7).

Step 4 The predefined elicitation method assigned to this position within the path was displayed. The procedure and additional content shown had been identical to those in Step 2.

Step 5 The questionnaire corresponding to the elicitation method from Step 4 was then displayed, using the same procedure as in Step 3.

Step 6 Based on the data collected from the elicitation methods in Steps 2 and 4, a preference profile for the six first-level categories was calculated in real-time. Participants were then asked to decide which of the two profiles aligned better with their interests (Figure A.8).

Step 7 In the final step, participants were thanked for their participation and had the opportunity to leave a comment. Any comments submitted were also stored in the corresponding dataset within the database table *xpo-elicitation-user*.

The integrated questionnaire used in Steps 3 and 5 consisted of five propositions measuring *choice difficulty* and *ease of use*. *Choice difficulty* included two propositions aimed at assessing the decision-making process for individual items, while *ease of use* comprised three propositions in which participants evaluated the overall process of the prior completed method. The measures and propositions were adopted or conceptually adapted from previous research on UI design, RSs, and choice overload [CT11] [KWG⁺12].

Measure	Propositions
Choice difficulty	It was easy for me to decide on the respective answer options. It didn't take long to decide on the respective answer options.
Ease of use	I found the whole approach to be very user-friendly. It was easy to understand how to use it. The process was time-consuming.

The responses to this questionnaire helped assess participants' experiences regarding choice difficulty and ease of use for each method. Furthermore, they contributed to answering RQ1.

3.3 Image Selection and Item Pool

The corporate partner provided a list of 109 potential categories of leisure time activities that could be presented to study participants. These categories had previously been grouped according to their overarching thematic areas, resulting in six first-order categories. Depending on the diversity of given leisure time activities, these first-order categories were assigned either one or two levels of sub-categories. This process resulted in the following first-order categories, organized based on their number of sub-categories. As the survey was conducted with German-speaking participants, the German name is written in brackets after the respective category name in the following list.

- **Sports** ("Sport"). With 53 sub-categories, *Sports* represented the largest grouping, including activities such as climbing, football, windsurfing, or skiing.
- **Entertainment** ("Unterhaltung"). The second largest category addressed entertainment-related items, including options such as karaoke, escape rooms, or pubquiz. 29 sub-categories were assigned to this category.
- **Cultural activities** ("Kultur"). This category consisted of eleven sub-categories and focused predominately on sightseeing, museums, exhibitions, or concerts.
- **Animal-related activities** ("Tiere"). Consisting of 5 sub-categories, this type of activities included options like visiting a zoo, a circus, or hiking with alpacas.

- **Course** ("Kurs"). This category included structured learning or skill-building activities, like dance classes, doing handicraft, or learning a language, organized into three sub-categories.
- **Relaxation** ("Entspannung"). This category encompassed activities aimed at reducing stress and promoting well-being, like wellness or meditation, divided into two sub-categories.

Based on this categorization and the characteristics of the investigated elicitation methods, a well-suited number of queries required for each method had to be determined. To obtain meaningful results, it was necessary to find a balance between collecting a sufficiently large amount of data per participant while preventing cognitive overload. Prior research had indicated that an overly-extensive questionnaire could lead to loss of attentiveness, which can both lead to biased survey results and dropout [KWG⁺12] [SLL⁺13].

Therefore, the number of steps and items per elicitation method was defined based on a balance between data quality, cognitive load, and user experience. For Swipe method and Rating method, 18 steps were chosen due to their lower complexity and faster interaction pace. In contrast, the Two-Items method and Four-Items method require more complex comparative decisions. To prevent cognitive overload and maintain engagement, the number of steps was reduced to 10 and 12 respectively, while still ensuring a comparable number of total evaluated items (20 and 48). To allow for meaningful profile generation and avoid repetitiveness, at least three distinct items per category were selected. This selection followed a structured, collaborative process aimed at ensuring that the visual representations of each category were both intuitive and consistent. This step was particularly important, as the images served as the primary content of the items presented to participants.

Initially, a pool of 96 images was pre-selected and assigned to the respective categories. Each category was initially represented by four images. To validate and refine this initial image assignment, a workshop was conducted with representatives from the corporate partner. During this session, the participants were asked to evaluate and rank the images within each category according to how well each image represented the category's content. In a second evaluation step, the two highest-ranked images per category, based on the number of individual votes, were reviewed once again. The participants then had the task of reaching consensus on the final image selection. This was to ensure that the selected images not only matched the semantics of the category, but also met the expectations of the users. The resulting final set of images was used in the elicitation items throughout the survey.

3.4 Preference Profile Calculation

The preference profile is a core component in RSs, serving as a representation of a user's interests, behaviors, and characteristics. In a tag-based RS, the preference profile is

constructed using tags associated with items that users interact with. This profile enables the system to provide personalized recommendations by matching user preferences with available items. This section details the calculation of the preference profile for each of the four elicitation methods.

In this study, the preference profile was represented as a vector with dimensions 1×6 , corresponding to the six first-order categories. Each participant's ratings from the elicitation method were aggregated and converted into a category-specific weighting. These individual category scores were then normalized to create a consistent vector representation. The resulting normalized 1×6 vector thus reflected the relative strength of a participant's preferences across the six first-order categories, facilitating a direct and meaningful comparison between preference profiles derived from the four elicitation methods.

For the preference profile calculation of each elicitation method the following was defined:

- U was a set of users
- I was a set of items
- T was a set of tags (i.e. animals, course, culture, entertainment, relaxation, sport, representing the first-order categories)
- Each item $i \in I$ was assigned exactly one tag $t(i) \in T$

3.4.1 Swipe method

The calculation of the preference profile for Swipe method proceeded as follows: A user u provided ratings $r_{u,i} \in \{-1, 1\}$ for those items i the user had rated. First, the raw tag-score for user u on tag t was calculated by summing all of the user's item-ratings corresponding to that tag:

$$x_u(t) = \sum_{\substack{i \in I \\ t \in T_i}} r_{u,i}$$

Next, the maximum tag-score was identified:

$$Max_u = \max_{t \in T} x_u(t)$$

The normalized preference profile P_u then resulted in:

$$P_u = \begin{cases} 0, & \text{if } \max_{t \in T} x_u(t) < 0, \\ \frac{1}{Max_u} (x_u(t_1), x_u(t_2), \dots, x_u(t_{|T|})), & \text{otherwise.} \end{cases}$$

3.4.2 Rating method

The calculation of the preference profile for Rating method proceeded as follows: A user u provided ratings $r_{u,i} \in \{1, 2, 3, 4, 5\}$ for those items i the user had rated. First, the raw tag-score for user u on tag t was calculated by summing all of the user's item-ratings corresponding to that tag:

$$x_u(t) = \sum_{\substack{i \in I \\ t \in T_i}} r_{u,i}$$

Next, the maximum possible accumulated rating across tags was identified:

$$Max_u = \max_{t \in T} c_t \times 5$$

where c_t was the number of items rated by user u with tag t .

The normalized preference profile P_u then resulted in:

$$P_u = \frac{1}{Max_u} (x_u(t_1), x_u(t_2), \dots, x_u(t_{|T|}))$$

3.4.3 Two-Items method

For the Two-Items method the calculation proceeded as follows: A user u provided ratings $r_{u,i} \in \{0, 1\}$ for those items i the user had rated. First, the raw tag-score for user u on tag t was calculated by summing all of the individual user's item-ratings corresponding to that tag:

$$x_u(t) = \sum_{\substack{i \in I \\ t \in T_i}} r_{u,i}$$

Next, the maximum tag-score was identified:

$$Max_u = \max_{t \in T} x_u(t)$$

The normalized preference profile P_u then resulted in:

$$P_u = \frac{1}{Max_u} (x_u(t_1), x_u(t_2), \dots, x_u(t_{|T|}))$$

3.4.4 Four-Items method

For the Four-Items method the calculation proceeded as follows: A user u provided ratings $r_{u,i} \in \{-1, 0, 1\}$ for those items i the user had rated. First, the raw tag-score for user u on tag t was calculated by summing all of the individual user's item-ratings corresponding to that tag:

$$x_u(t) = \sum_{\substack{i \in I \\ t \in T_i}} r_{u,i}$$

NowNext, the minimum tag-score was identified:

$$Min_u = \min_{t \in T} x_u(t)$$

If $Min_u < 0$ the vector was shifted to the nonnegative range and the maximum tag-score was identified:

$$x'_u(t) = x_u(t) + |Min_u|$$

$$Max_u = \max_{t \in T} x'_u(t)$$

The normalized preference profile P_u then resulted in:

$$P_u = \frac{1}{Max_u} (x'_u(t_1), x'_u(t_2), \dots, x'_u(t_{|T|}))$$

If $Min_u > 0$ the maximum tag-score is identified:

$$Max_u = \max_{t \in T} x_u(t)$$

The normalized preference profile P_u then resulted in:

$$P_u = \frac{1}{Max_u} (x_u(t_1), x_u(t_2), \dots, x_u(t_{|T|}))$$

3.4.5 Advanced Weighting Techniques in Tag-Based RS

Several optimizations can enhance the calculation of preference profiles in tag-based RSs. Temporal weighting assigns a higher weight to tags associated with more recent interactions, reflecting the assumption that recent activities are more indicating on current preferences [LKS⁺14]. Contextual weighting involves weighting tags based on the context in which they are used, such as time of day, location, or specific activities, thereby increasing the relevance of recommendations by considering situational factors [GT11]. Tag hierarchies and ontologies utilize hierarchical structures or ontologies to understand the relationships between tags. This enables more nuanced recommendations, as the system can recommend items that are semantically related to the user's preferences, even if they do not have the exact same tags [ZDN⁺08].

Despite the potential advantages of these optimizations, a simpler calculation method was used in this study for several reasons. A simpler model is easier to implement, interpret, and validate, especially in the initial stages of research, and provides a clear baseline for understanding the effectiveness of the four elicitation methods without the additional complexity of advanced weighting schemes. The study, by its design, was not intended to collect data required for more complex weighting schemes, such as detailed temporal or contextual information for each interaction. The primary goal was to compare the effectiveness of the elicitation methods investigated. Introducing additional complexity into the profile calculation could have obscured the differences resulting from the methods themselves.

3.5 Evaluation and Participant Dimensions

This section outlines the criteria used to measure method performance, i.e. completion rate, time efficiency, usability, and profile fit. Additionally, relevant participant character-

istics were considered to enable differentiated analysis and to identify potential patterns or biases in the results.

3.5.1 Definition of Criteria for Evaluation

To quantitatively assess the usability of each elicitation method, the following criteria were defined based on established principles from usability engineering and decision science:

- **Completion rate:** The ratio of completed surveys for each method. As highlighted by Sauro & Lewis [SL16], completion rate serves as a fundamental usability metric.
- **Time efficiency:** The duration required to complete the respective elicitation method. Following to Frøkjær et al. [FHH00], task completion time is a core indicator of usability efficiency.
- **Usability:** How user-friendly the respective elicitation method was perceived. According to Winter and colleagues [WWD08], perceived usability reflects how effectively and efficiently users can interact with a method to achieve their goals.
- **Profile Fit:** Profile Fit is assessed subjectively by participants themselves, based on which of the resulting profiles represents their preferences more accurately.

Overall, these dimensions provide a comprehensive and user-centered basis for comparing the practical applicability and perceived effectiveness of each elicitation method.

3.5.2 Definition of Criteria for Participants

To ensure meaningful interpretation of the results and to identify potential influencing factors, the selection of participant characteristics was grounded in established research practices:

- **Gender:** Gender was included to examine potential gender-specific differences in preferences and interaction patterns. Prior research has shown that gender can influence technology acceptance and user behavior [VM00].
- **Age Range:** Age groups were defined to capture different life stages and levels of digital experience. As highlighted by Czaja et al. [CCF⁺06], age has been shown to affect cognitive processing and technology use behavior.
- **Frequency of Activities:** Regular engagement in leisure activities was considered a relevant factor. This context aids in interpreting how lifestyle and domain knowledge may influence preference formation [KZFM09].

3.6 Data Collection

Participation in the study was voluntary and participants were recruited via snowball system, mainly through private messages, social media channels and email mailings. Capturing their ratings was crucial to obtain meaningful results in the subsequent data evaluation process. To maximize data collection, ratings were not stored locally in the participant's browser and submitted collectively at the end. Instead, ratings were saved in the database at each transition to the next step in the survey process. This approach would have allowed for an analysis of when and at which step a session was canceled. However, the specific reasons for dropout could not be determined. Potential explanations include:

- Decision by the participant to stop due to lack of interest in the method or topic [RGR24].
- An unintentional interruption, such as a phone call or other distraction, preventing the participant from re-engaging, or an extended period of inactivity leading to the session timing out.
- A network connectivity issue between the participant's device and the server.

To ensure a solid foundation for subsequent data analysis, both the interaction data and questionnaire responses were systematically stored and structured.

3.6.1 Data Storage of Elicitation Method

For each item in a given elicitation process, a corresponding record was created in the database. This resulted in one record per item for the Swipe method and Rating method, two records for the Two-Items method, and four records for the Four-Items method per step.

The following data points were stored in the database table *xpo-elicitation-item-answer*:

- **User:** The identifier for the participant.
- **Item:** The specific item being evaluated.
- **Rating:** The participant's assigned evaluation.
- **Timestamp:** The exact time the record was saved.
- **Survey Path:** The predefined path of the survey session.
- **Displayed Items:** The set of items presented in the given step.

For the Two-Items method and Four-Items method, an additional timestamp was recorded for the moment participants clicked the "heart" or "X" symbol. In Four-Items method, if participants initially liked an item, but later changed their selection, the timestamp of the first liked item was still stored. These data would enable an analysis of which item pairs participants found difficult to choose between, and which item they finally selected.

3.6.2 Data Storage of Questionnaire

The responses from the questionnaires were stored uniformly across all elicitation methods.

The following data points were stored in the database table *xpo-elicitation-question-answer*:

- **User:** The identifier for the participant.
- **Proposition:** The specific proposition being answered.
- **Rating:** The participant's response or evaluation.
- **Timestamp:** The exact time the record was saved.
- **Survey Path:** The predefined path of the survey session.
- **Questionnaire Position:** The step within the survey where the questionnaire was presented.

The structured collection and storage of both interaction data and questionnaire responses formed the technical basis for the subsequent system implementation. The following Chapter 4 presents the design considerations and concrete realization of the elicitation methods described above.

Design and Implementation

4.1 Artefact Architecture and System Components

The implemented artefact is a browser-based prototype developed from scratch to simulate and evaluate different visual preference elicitation methods in the context of a mobile leisure RS. The primary goal was to create an accessible and testable application while ensuring alignment with real-world use case.

4.1.1 Backend

The backend of the prototype was implemented using PHP in version 7.4, a widely used scripting language in web development. To support the separation of logic and presentation layers, the Smarty template engine in version 3.1 was integrated. This approach contributed to a clearer code structure and facilitated the maintainability of the application.

4.1.2 Frontend

The client-side Graphical User Interface (GUI) was developed using HTML5, the CSS framework Bootstrap in version 4.6, and the JavaScript library jQuery in version 3.6. HTML5 provided the structure for the web pages, while Bootstrap enabled a responsive layout across varying screen sizes. jQuery was used to implement interactive elements and dynamic content updates.

4.1.3 Database

The ratings submitted by participants, along with questionnaire responses and demographic information were stored in a relational database. For this purpose, MariaDB in version 10.11 was used, providing a reliable open source solution for managing structured

data. The relational database model allowed for efficient querying and management of the collected data, supporting the analysis and evaluation phases of the study.

4.1.4 Communication Layer

The communication between frontend and backend was handled via AJAX requests, allowing asynchronous data exchange during the survey process. Data transfer in JavaScript Object Notation (JSON) format enabled an efficient loading of elicitation items and saving of user responses between survey steps without full page reloads. This setup contributed to a more fluid user experience during the elicitation process.

4.2 User Interface Design

The interface was intentionally kept minimalistic to retain participants' focus. This approach aimed to prevent visual overload, minimize distractions, and ensure accessibility for participants. For all survey steps, a light gray (#EDED) background and black (#000000) text color were used.¹ The instructional pop-up for each elicitation method was displayed with black (#000000) text color on a white (FFFFFF) background.

Each item was presented as a tile containing a representative image of a leisure activity category. A white text label with the category name was overlaid at the bottom of the image, placed on a transparent gradient in the primary color of the corporate partner's Corporate Identity (CI). This design ensured text readability, even if bright image areas were present. This design decision is grounded in cognitive load theory and Human-Computer Interaction (HCI) principles. According to Oviatt [Ovi06], reducing visual complexity enhances task focus and minimizes the cognitive effort required to perform decisions, which is important for mobile environments where screen space and attention are limited.

An alternative layout in which the category name and a short description were placed below the image, was considered but ultimately discarded. Positioning the text outside the image would have required reducing the image size, limiting its visual impact. The inclusion of a short description was also discarded to prevent any potential bias.

Below the image, interactive controls were displayed, allowing participants to evaluate the current item. The interface was designed to reserve sufficient space at the bottom of the screen, ensuring that the interactive controls remained fully visible without the need for vertical scrolling. The survey was fully responsive to ensure optimal presentation across different devices, with images dynamically scaling to fit the available screen space without distortion. An exception to this layout was Four-Items method, where four images were shown in a 2×2 grid. In this layout, the controls of the presented items in the second row were positioned outside the initially visible area, requiring participants to scroll vertically to access them. However, since the Four-Items method inherently

¹The values in parentheses refer to hexadecimal color codes.

required vertical scrolling to view and assess all items, this adjustment was considered necessary and acceptable.

Consistency in layout and control positioning supports more efficient decision-making and accelerates onboarding processes [MP09].

4.3 Implementation of Elicitation Methods

The following subsections describe the user interface behavior and interaction mechanisms specific to each method. Particular attention was paid to ensuring an intuitive and consistent user experience across all methods while aligning the interaction design with the underlying logic of each elicitation approach.

4.3.1 Swipe Method

The Swipe method allowed participants to choose between two predefined symbols: a "heart" symbol for a positive rating and an "X" symbol for a negative rating (Figure A.2). To encourage participants to complete the survey, a playful gesture-based interaction was incorporated for submitting ratings. Swiping left assigned a negative rating, while swiping right marked the item as positive.

As participants performed the swiping gesture, a corresponding "X" or "heart" symbol appeared on the item, providing direct visual feedback aligned with the chosen direction. This feature allowed participants to reconsider their choice by returning the item to its original position before finalizing their decision (Figure A.3). To confirm their decision using the swipe mechanism, participants had to release the item while in motion to the left or right side. Once participants made a selection using the controls, the decisions could not be reversed, and the next item for evaluation was presented.

4.3.2 Rating Method

The 1 to 5 point scale (1 = lowest rating; 5 = highest rating) of the Rating method was visually represented using star symbols. Clicking on a star filled in the selected star along with all stars to its left, ensuring that participants could clearly recognize their chosen rating. They were able to revise their decision at any time by selecting a different star. Once a rating had been made, a "Continue" button appeared, allowing respondents to proceed to the next item (Figure A.1).

4.3.3 Two-Items Method

In the Two-Items method, the items to be evaluated were displayed side by side, with a "heart" symbol beneath each item allowing participants to indicate their preference with a click. Initially, the color of the "heart" symbol was grey, indicating an inactive state. If no selection was made within three seconds, a pulsating animation was triggered around the "heart" symbols to draw attention to the interaction mechanism.

Once a decision had been made and one of the two items was selected, the "heart" symbol of the chosen item turned pink, while the "heart" symbol of the other item was hidden. A "Continue" button then appeared below, enabling participants to proceed to the next item pair for evaluation (Figure A.4). By clicking the "heart" symbol of the selected item again, the view returned to its initial state, allowing participants to revoke their decision.

4.3.4 Four-Items Method

In the Four-Items method, the items were displayed in a 2×2 grid. Initially, a "heart" symbol appeared beneath each of the four items. As in the Two-Items method, the "heart" symbols were displayed in gray to indicate an inactive state. Once an activity was selected, the color of the symbol associated with the chosen item changed to pink, visually confirming the participant's choice (Figure A.5).

Subsequently, the "heart" symbols under the remaining items were replaced with gray "X" symbols, prompting participants to select the category they liked the least. The "X" symbol of the least preferred category turned red as soon as it was clicked, while the controls for the remaining, unrated items were hidden.

As in the Two-Items method, if no interaction occurred within three seconds, a pulsating animation was triggered around the preference selection symbols to guide participants on how to make a choice. After both a positive and a negative selection were confirmed, a "Continue" button appeared, allowing participants to proceed to the next set of categories for evaluation (Figure A.6). Participants also had the option to undo their selection by clicking the symbol of the most recently chosen item again, just as in the Two-Items method.

4.4 Item Selection and Control Logic

This section outlines the underlying logic used to determine which items were displayed to participants during the elicitation process. It describes how item selection was dynamically adapted based on user input and method-specific requirements, ensuring a balanced and meaningful distribution of content across all survey paths.

4.4.1 Swipe and Rating Method

Since both the methods require the evaluation of only a single item per step, the same algorithmic approach was applied for the selection and sequencing of items. The selection process began with first-order categories, such as *Sports* or *Culture*. Participants were shown items from these top-level categories they had not yet evaluated. The system ensured that each first-order category was presented at least once and adaptively avoided repeating previously rated items.

The transition from first-order to second-order items was not strictly tied to a fixed step (e.g., the seventh item), but was driven by the structure of the participants' responses.

Once all first-order categories had been evaluated, the algorithm moved forward by selecting subcategories from the preferred areas. For example, if participants had expressed interest in *Sports*, they were subsequently shown subcategories such as *Motor Sports* or *Water Sports*. If no first-order category had been rated positively, the system instead selected a subcategory from a previously disliked category. To maintain balance and ensure diversity, the algorithm kept track of how many subcategories had already been shown under each first-order category and prioritized those with fewer exposures.

Additionally, the selection process included a control mechanism based on a modulo operation. Specifically, after every fourth positively rated second-order item, the system inserted a subcategory item from a first-order category the participant had previously rated negatively. This step served to test the consistency of preferences and to help identify careless or inattentive response patterns. Within this process, the algorithm ensured that the selected second-order item came from a category with the fewest prior evaluations, promoting a well-distributed and balanced presentation across the elicitation process.

4.4.2 Two-Items Method

In each step, participants assigned to the Two-Items method were presented with two choices and were asked to identify their preferred option. The unselected item remained unassessed and was considered neutral for that step. The algorithm followed a structured sequence to ensure a balanced preference elicitation process.

Initially, two items were selected from categories within the first order. In the second step, two additional items were drawn from first-order categories that had not been chosen in Step 1. This process was repeated once more, ensuring that all first-order categories were considered. For instance, if participants were shown items from the categories *Entertainment* and *Animal* in the first step, these categories did not reappear in Steps 2 and 3.

Following the initial selection phase, the algorithm proceeded to select two items from categories that had not been preferred in the earlier rounds. Subsequently, two more items were chosen from the remaining categories that had not yet been favored. As the process continued, the algorithm revisited those categories that had previously been selected as preferred. Two items were drawn from these categories to reinforce the participants' initial preferences.

In Step 7, one item was selected from a category that had been preferred in earlier steps, but not in the immediate prior round, while the second item was drawn from a category preferred in Step 6. Further refinement took place by selecting two items from categories that had previously been preferred, but had not been selected in Steps 6 and 7.

In the final stages, items were chosen from categories that had been prioritized in the middle of the process. First, items from categories preferred in Steps 4 and 5 were displayed, followed by items from categories that had not been favored in those same

steps. For example, if participants had selected "Visiting the Zoo" (*Animal* category) over "Attending a Theater Performance" (*Cultural* category) in Step 4, another item from the *Animal* category was shown in Step 9, while an item from the *Cultural* category appeared in Step 10.

In summary, this approach ensured a balanced representation of both preferred and non-preferred categories, thereby avoiding monotonous answer patterns and reducing the risk of losing participants' attention. As the process advanced, the selection criteria became increasingly specific while still guaranteeing the holistic inclusion of categories, which prevented the over-representation of certain topics.

4.4.3 Four-Items Method

Participants assigned to this method were asked to evaluate four items at once and to choose both their most preferred and least preferred option. The two remaining items in each step were considered neutral and did not contribute to the preference assessment.

In contrast to the other three approaches, the Four-Items method did not adapt the selection of displayed items based on participants' previous preferences. Instead, the items, referred to as tuples, were generated in advance for each step using a randomized sampling process. These tuples were created according to a set of predefined criteria:

- No two tuple have the same four items.
- No two elements within a tuple were identical.
- Each entry in the item list appeared in approximately the same number of tuples.
- Each pair of entries appeared together in approximately the same number of tuples.

According to [KM16], generating between 1.5 to 2 times the total number of elements (denoted as N) was sufficient to achieve reliable results. Given that the selection process involved six first-level categories, this approach required the creation of 12 tuples.

The algorithm for generating these tuples followed a structured sequence of steps:

1. The six first-order categories were selected, and a list containing their category IDs was created.
2. A loop was initiated to repeat steps 3 to 7 one hundred times.
3. The list of category IDs was shuffled to generate a randomized sequence of elements.
4. The elements were grouped into tuples of four. If there were not enough remaining elements to form a complete tuple, additional random elements were added.
5. The frequency of item pairings within the created tuples was calculated.

6. The standard deviation of the frequency distribution of these pairs was computed to assess the balance of the generated combinations.
7. The combination with the lowest standard deviation was selected, ensuring an even distribution of pair occurrences.

Once the 12 tuples had been generated, the algorithm iterated through them, selecting second-order category items corresponding to the overarching first-order categories in each tuple. This ensured that different items were shown in each round of the evaluation process.

Additionally, the selection algorithm was designed to distribute items evenly across all tuples in order to minimize repetition. The generated tuples were then stored in the dataset associated with each participant. As the survey progressed, the corresponding tuple for each step was retrieved and presented to the participant for evaluation.

4.5 Iterative Refinement of the Prototype

During the implementation phase, the functionality and usability of the prototype were continuously evaluated and refined through two distinct test cycles, in line with the iterative nature of the Design Cycle in the DSR framework.

The first test cycle involved representatives of the corporate partner who interacted with a preliminary version of the artefact. Their feedback led to initial adjustments to the user interface and interaction logic. Specifically, improvements were made to the clarity of interface elements, the placement of interaction controls, and the explanatory text used in pop-ups.

In the second refinement round, a broader test group consisting of twelve participants provided detailed feedback regarding usability and interaction design. The most important findings included:

- Inconsistent recognition of swipe gestures in the Swipe method.
- Uncertainty about the number of remaining steps.
- Lack of visible interaction instructions in the Four-Items method, leading to confusion about the intended functionality.
- Late discovery of the help icon; users preferred that the instruction pop-up be displayed immediately upon entering a method.
- In rare cases, participants who consistently gave negative ratings in the Swipe method experienced a halt in item delivery after several steps due to the exhaustion of valid options.
- Difficulties in reading the white text label on light-colored image backgrounds.

- Challenges in interpreting the radar chart on the profile comparison screen.

These insights were systematically analyzed and transformed into concrete improvements to the prototype:

- Swipe recognition was improved by increasing gesture sensitivity and enlarging the interactive touch areas.
- A visual step counter (e.g., "6/18") was introduced across all methods to enhance user orientation.
- A pulsating animation was implemented in the Two-Items method and Four-Items method to guide users in the absence of immediate interaction.
- The instruction pop-up was automatically triggered at the beginning of each elicitation method.
- Error-handling mechanisms were added to prevent cases where no further items could be displayed.
- Visual contrast was improved by applying a gradient overlay to image tiles, ensuring better readability of text labels.
- The profile comparison screen was extended with an additional ranked list to support users unfamiliar with radar chart representations.

Based on these iterative refinement cycles, the prototype was gradually improved in order to provide a stable, user-friendly and functionally complete version for the subsequent evaluation phase. This iterative approach ensured that the prototype not only met the technical requirements, but also enhanced its usability and effectiveness in capturing user preferences.

Evaluation

5.1 Participants

The survey was conducted over a two-month period, beginning on March 9, 2023. During this time, a total of 654 participants initiated one of the three survey modes. As participation was voluntary, individuals had the option to discontinue at any time. Of those who started the survey, 486 participants completed the full process. These formed the baseline for further analyses.

Since the survey was developed and conducted as a browser-based application, it was accessible on any device type. Table 5.1 presents the device types used by participants who completed the study.

Device types	Participants
Smartphone	382
Desktop	91
Tablet	9
Phablet	4

Table 5.1: Participants grouped by device types.

Since this study focused on the corporate partner’s mobile app, which was optimized exclusively for smartphones, only data from participants who completed the survey using a smartphone were considered for further analysis. After filtering the data to include only smartphone users, 525 participants started the survey, and 382 successfully completed the process, resulting in a completion rate of 72.76%. Table 5.2 presents the distribution of participants across the three survey modes.

Survey path	Participants
Comparison of Swipe method and Four-Items method	128
Comparison of Swipe method and Rating method	128
Comparison of Four-Items method and Two-Items method	126

Table 5.2: Number of participants who completed respectiv survey path.

The number of participants was nearly evenly distributed, ensuring a balanced dataset and the algorithm for assigning a survey path to participants, as described in Section 3.2, enabled a meaningful and reliable comparison of the different survey modes.

The Participants were not monitored while conducting the survey. Their attention and motivation to complete the elicitation process properly were neither recorded nor observed. Additionally, all responses, including demographic information such as age group, gender, and frequency of leisure activities, were provided voluntarily.

As shown in Table 5.3, six (1.57%) participants chose not to disclose their age. Among those who provided this information, 296 (77.49%) reported being in the 1–24 age group, followed by 75 (19.63%) in the 25–34 age group, and four (1.05%) in the 35–44 age group. No participants selected the 45–54 or 55–64 age groups; however, one (0.26%) participant indicated to be 65 years or older.

Age range	Participants	
Unknown	6	1.57%
1-24	296	77.49%
25-34	75	19.63%
35-44	4	1.05%
45-54	0	0.00%
55-64	0	0.00%
65+	1	0.26%

Table 5.3: Participants grouped by age ranges.

The majority of participants identified as female (240 participants, 62.83%), while 128 (33.51%) participants indicated they were male. Additionally, six (1.57%) participants described themselves as non-binary, and eight (2.09%) participants chose not to disclose their gender. This distribution is presented in table 5.4.

The frequency with which participants engage in leisure activities is presented in Table 5.5. The largest group, 238 (62.30%) participants, reported going out two to four times per

Gender	Participants	
Unknown	8	2.09%
Female	240	62.83%
Male	128	33.51%
Non-binary	6	1.57%

Table 5.4: Participants grouped by gender.

month, while 84 (21.99%) participants indicated they are active four times per week or more. Additionally, 41 (10.73%) participants engage in activities at most once per month, and 19 (4.97%) participants did not provide a response.

Frequency of activities	Participants	
Unknown	19	4.97%
Four times a week or more	84	21.99%
Two to four times a month	238	62.30%
Once a month or less often	41	10.73%

Table 5.5: Participants grouped by frequency of activities.

5.2 Completion Rate

The measure *Completion Rate* represents the percentage of participants who started and completed the respective evaluated method in their assigned survey path. A low completion rate may indicate difficulties with usability, engagement, or cognitive load associated with the method. Table 5.6 presents the completion rates of the four elicitation methods. The Two-Items method had the lowest completion rate at 66.09%, followed by the Rating method with 70.62%. In contrast, the Swipe method had the highest completion rate at 82.91%, while the Four-Items method had a completion rate of 80.17%. A Chi-square test of independence was conducted to examine the relationship between the elicitation method and the completion rate. The results showed a significant correlation between the method used and the likelihood of completing the elicitation process ($\chi^2(3) = 24.743, p < 0.0001$).

5.3 Time Efficiency

The measure *Time Efficiency* represents the average time, in seconds, that participants required to complete the respective elicitation process. This metric provides insight into

	Completion Rate
Swipe method	82.91%
Rating method	70.62%
Two-items method	66.09%
Four-items method	80.17%

Table 5.6: Completion rate in percent per elicitation method.

the efficiency and cognitive load of the different methods. Table 5.7 presents the mean completion time across all participants and further segments the results by gender, age group, and frequency of leisure activities.

Overall Completion Time

On average, the Two-Items method was the fastest to complete (45.23 seconds), followed by the Swipe method (58.62 seconds) and the Rating method (68.34 seconds). The Four-Items method required significantly more time (148.07 seconds). The boxplot in Figure 5.1 shows the differences in completion times across the four elicitation methods. The Four-items method required the most time, with a median completion time above 100 seconds and a wide distribution, including several outliers exceeding 250 seconds. The Two-Items method, Swipe method, and Rating method had significantly shorter completion times, than the Four-Items method with medians below 55 seconds. Among these, the Two-Items method, as the fastest, is also displaying the least variance.

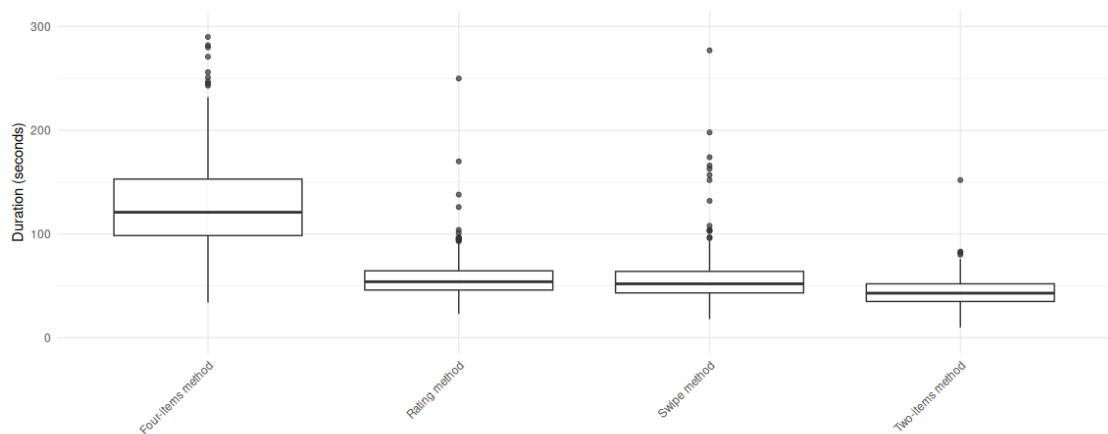


Figure 5.1: Duration distributions of completed elicitation methods.

These results indicate that task complexity directly influences response time, as the Four-Items method requires participants to select both a most and least preferred item, making

	Swipe	Rating	Two-items	Four-items
All	58.62	68.34	45.23	148.07
Gender				
Unknown	54.31	53.86	38.50	243.57
Female	58.89	58.44	45.83	135.25
Male	61.67	66.57	44.64	141.05
Non-binary	50.40	506.00	-	101.25
Age group				
Unknown	53.91	48.14	38.50	253.29
1-24	59.01	62.88	44.87	135.74
25-34	61.58	56.76	47.67	142.04
35-44	82.33	532.00	-	93.00
45-54	-	-	-	-
55-64	-	-	-	-
65+	55.00	-	-	68.00
Frequency of activities				
Unknown	54.30	53.92	44.00	230.14
Four times a week or more	53.91	58.50	45.54	132.50
Two to four times a month	62.22	62.71	44.92	136.67
Once a month or less often	57.89	138.50	47.50	139.32

Table 5.7: Average duration (in seconds) summary across four elicitation methods, presented for all participants and further segmented by gender, age group, and frequency of activities.

it more cognitively demanding than the simpler Two-Items method, where only one choice is required. The presence of outliers in the Four-Items method also suggests that some participants struggled with this method, leading to significantly longer completion times.

Completion Time by Gender

The time required to complete the elicitation process varied slightly by gender. Male participants generally took longer than female participants in all methods except the Two-Items method. For example, males needed more time in the Swipe method (61.67 seconds) compared to females (58.89 seconds) and in the Rating method (66.57 seconds and 58.44 seconds, respectively) as well. Non-binary participants showed the highest variance, particularly in the Rating method (506.00 seconds). This is also confirmed in the boxplot in Figure 5.2 where the Four-Items method consistently required the most time across all gender groups and the variance in the Rating method was noticeably

higher among non-binary participants.

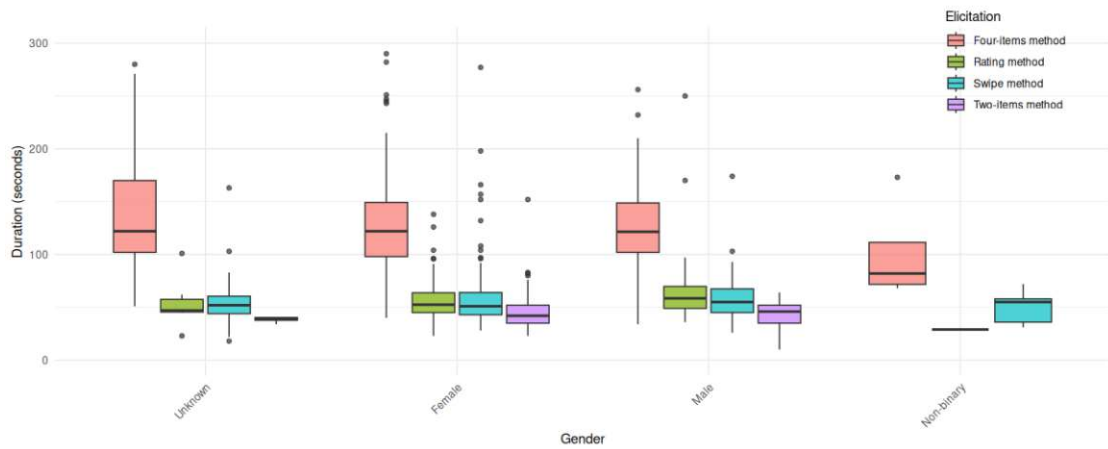


Figure 5.2: Duration distributions of completed elicitation methods of different genders.

These results may suggest that certain users found specific methods more challenging, though gender itself does not appear to have a strong influence on completion time. But it is more likely due to the small sample size of only six participants in the Non-binary gender group, making the mean more vulnerable to the influence of outliers.

Completion Time by Age Group

Age-related differences in completion time are also evident in Table 5.7. Participants aged 1-24 years and 25-34 years showed similar performance across methods, with durations ranging between 44.87 and 62.88 seconds for the first three methods and twice the time for the Four-Items method. In contrast, the 35-44 age group completed the Four-Items method on average in 93 seconds, much faster than the other groups, but exhibited a significantly higher completion time for the Rating method (532.00 seconds). This was likely due to the small sample size making the mean highly sensitive to individual variations. Finally, the 65+ participant completed the tasks (Swipe method in 55.00 seconds and the Four-Items method in 68.00 seconds) in times comparable to younger participants, suggesting that age did not strongly impact completion time. These findings are further visualized in the boxplot presented in Figure 5.3.

Completion Time by Frequency of Activities

Participants who engage in leisure activities more frequently tended to complete the survey slightly faster across most methods. Those who reported going out four times a week or more completed the elicitation tasks in 45.54 to 58.50 seconds for the first three methods, while requiring 132.50 seconds on average for the Four-Items method. Similarly, participants who engage in activities two to four times a month showed slightly longer completion times, ranging from 44.92 to 62.71 seconds, with a Four-Items method

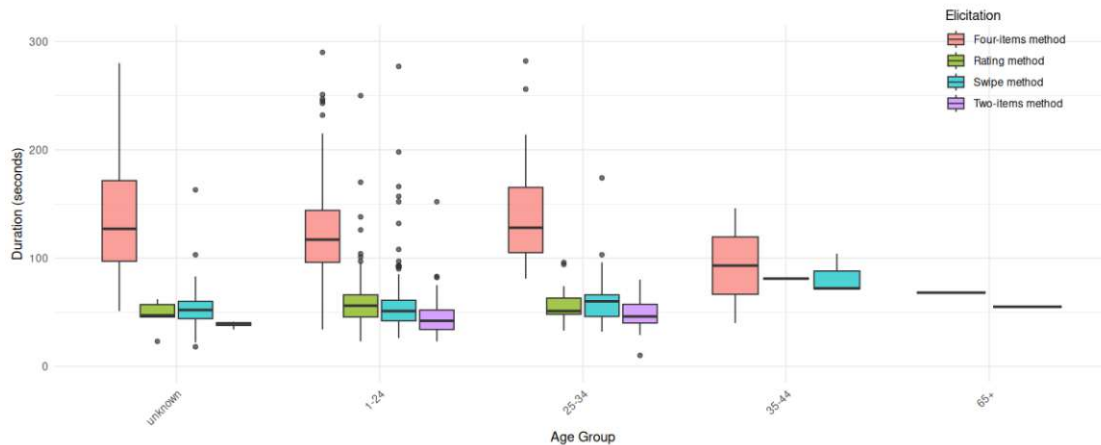


Figure 5.3: Duration distributions of completed elicitation methods of different age groups.

duration of 136.67 seconds. In contrast, participants who engage in activities once a month or less often took significantly longer to complete the Rating method (138.50 seconds), and their completion time for the Four-Items method (139.32 seconds) also exceeded that of the more active groups.

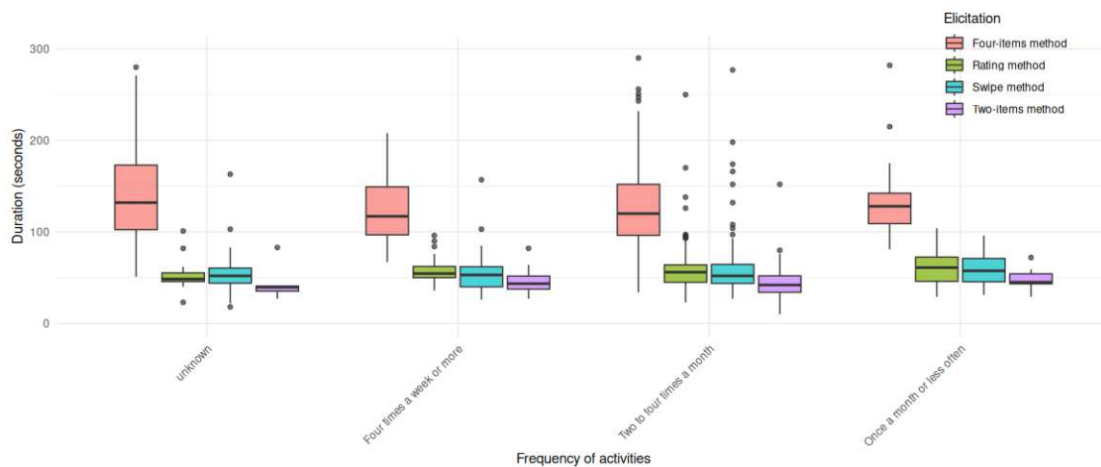


Figure 5.4: Duration distributions of completed elicitation methods of different activity frequencies.

Statistical Analysis of Method Efficiency

To examine whether the assumption of normality was met for the time efficiency results, a Shapiro-Wilk test was conducted separately for each elicitation method. Results revealed that the distribution of completion times significantly deviated from normality

for all methods (Swipe Method: $p < 0.0001$; Rating Method: $p < 0.0001$; Two-items Method: $p < 0.0001$; Four-items Method: $p < 0.0001$). Therefore, to further analyze the differences between the four elicitation methods, a Dunn test with Bonferroni correction was conducted following a significant Kruskal-Wallis test ($\chi^2(3) = 476.72, p < 0.0001$). The results are presented in Table 5.8.

	<i>Z – Value</i>	<i>p – Value</i> (Bonferroni)
Swipe and Four-items	17.97	< 0.0001
Swipe and Rating	1.36	0.5244
Swipe and Two-items	4.18	0.0001
Rating and Two-items	4.69	0.0001
Four-items and Rating	12.65	< 0.0001
Four-items and Two-items	17.75	< 0.0001

Table 5.8: Results of the Dunn test with Bonferroni correction for pairwise comparisons of the methods. Significant differences with $p < 0.05$.

The analysis reveals that the Four-Items method takes significantly longer than all other methods ($p < 0.0001$). The Two-Items method is significantly faster than the Rating method ($p = 0.0001$) and the Swipe method ($p = 0.0001$). No significant difference was found between the Rating method and Swipe method ($p = 0.5244$), indicating that these two methods are comparable in terms of completion time.

5.4 Usability

The measure *Usability* evaluates how participants perceived the ease of use of each elicitation method on a 1 to 5 Likert scale, where higher values indicate greater usability. Table 5.9 presents the average usability ratings for all participants and further segments the results by gender, age group, and frequency of activities.

Overall Usability Ratings

Across all participants, the Swipe method received the highest usability rating (4.56), followed closely by the Rating method (4.52) and the Two-Items method (4.47). The Four-Items method received the lowest usability rating (4.07), indicating that participants found it more complex or less intuitive compared to the other methods.

Usability Ratings by Gender

Across all gender groups, the Swipe method received the highest usability ratings, with female participants giving it a score of 4.64, male participants rating it 4.45, and non-binary participants assigning it 4.57. The Rating method also received relatively high usability scores, with females rating it 4.62, males 4.42, and non-binary participants significantly lower at 3.70. In comparison, the Two-Items method was rated similarly by

	Swipe	Rating	Two-items	Four-items
All	4.56	4.52	4.47	4.07
Gender				
Unknown	4.34	3.87	4.20	3.64
Female	4.64	4.62	4.48	4.11
Male	4.45	4.42	4.44	4.04
Non-binary	4.57	3.70	-	3.90
Age group				
Unknown	4.40	4.07	-	3.07
1-24	4.56	4.55	4.48	4.13
25-34	4.61	4.52	4.48	3.98
35-44	3.93	3.30	3.00	3.50
45-54	-	-	-	-
55-64	-	-	-	-
65+	4.00	-	-	1.40
Frequency of activities				
Unknown	4.51	4.20	4.80	4.07
Four times a week or more	4.54	4.53	4.34	3.99
Two to four times a month	4.60	4.58	4.48	4.13
Once a month or less often	4.39	4.32	4.60	3.97

Table 5.9: Average usability (1 to 5 Likert-Scale) summary across four elicitation methods, presented for all participants and further segmented by gender, age group, and frequency of activities.

females and males (4.48 and 4.44, respectively), while no usability rating was available for non-binary participants. The Four-Items method received the lowest usability scores across all gender groups, with females rating it 4.11, males 4.04, and non-binary participants 3.90.

Usability Ratings by Age Group

Analyzing usability ratings by age, the Swipe method consistently received high ratings, with participants aged 1-24 rating it 4.56, and those aged 25-34 giving it 4.61. In contrast, the 35-44 age group rated it considerably lower at 3.93, while the 65+ participant rated it 4.00. The Rating method followed a similar trend, with usability scores of 4.55 (1-24 years), 4.52 (25-34 years), and a notably lower score of 3.30 from the 35-44 group. In comparison, the Two-Items method had the same rating in the age groups 1-24 and 25-34, with a value of 4.48, and was lower in the 35-44 group, which rated it at 3.00. The Four-Items method consistently received the lowest ratings, particularly from the 65+

	<i>Z – Value</i>	<i>p – Value</i> (Bonferroni)
Swipe and Four-items	14.85	< 0.0001
Swipe and Rating	1.41	0.4782
Swipe and Two-items	2.88	0.0119
Rating and Two-items	1.29	0.5963
Four-items and Rating	10.73	< 0.0001
Four-items and Two-items	9.19	< 0.0001

Table 5.10: Results of the Dunn test with Bonferroni correction for pairwise comparisons of the methods. Significant differences with $p < 0.05$.

participant, who rated it 1.40.

Usability Ratings by Frequency of Activities

Participants who engage in leisure activities four times a week or more rated the Swipe method (4.54) and Rating method (4.53) highly, while the Two-Items method received a lower score of 4.34 and the Four-Items method was rated lowest at 3.99. Those who engage in activities two to four times a month showed similar trends, with the Swipe method and Rating method receiving the highest scores (4.60 and 4.58, respectively), while the Two-Items method was rated 4.48 and the Four-Items method 4.13. Participants who engage in activities once a month or less often rated the Two-Items method highest (4.60), followed by the Swipe method (4.39), with the Four-Items method receiving the lowest rating (3.97).

Statistical Analysis of Usability Ratings

To verify the assumption of normality for usability ratings, a Shapiro-Wilk test was conducted for each elicitation method separately. Results showed a significant deviation from normality across all groups (Swipe Method: $p < 0.0001$; Rating Method: $p < 0.0001$; Two-items Method: $p < 0.0001$; Four-items Method: $p < 0.0001$). Therefore, after a significant Kruskal-Wallis test ($\chi^2(3) = 252.46, p < 0.0001$), a Dunn test with Bonferroni correction was conducted, with results presented in Table 5.10, and identified statistically significant differences between several methods. The Swipe method had significantly higher usability ratings than the Four-Items method ($p < 0.0001$), indicating that participants found it easier to use. The Four-Items method was also rated significantly lower than the Rating method ($p < 0.0001$) and the Two-Items method ($p < 0.0001$). The Swipe method had significantly higher usability ratings than the Two-Items method ($p = 0.0119$). However, no significant difference was found between the Rating method and Two-Items method ($p = 0.5963$), nor between the Swipe method and Rating method ($p = 0.4782$), which may indicate that participants perceived them as similarly user-friendly.

5.5 Profile Fit

The *Profile Fit* measure evaluates how well the preference profile generated by the compared elicitation methods aligned with participants' interests. The results, presented in Tables 5.11, 5.12 and 5.13, summarize the percentage of participants who selected a particular method's generated preference profile as the one that represents their preferences best, depending on the assigned survey path. The data is further segmented by gender, age group, and frequency of leisure activities.

	Swipe	Four-items
All	42.97%	57.03%
Gender		
Unknown	0%	100%
Female	49.33%	50.67%
Male	31.11%	68.89%
Non-binary	100%	0%
Age group		
Unknown	0%	100%
1-24	44.55%	55.45%
25-34	36.36%	63.64%
35-44	100%	0%
45-54	-	-
55-64	-	-
65+	100%	0%
Frequency of activities		
Unknown	50%	50%
Four times a week or more	42.86%	57.14%
Two to four times a month	35.42%	64.58%
Once a month or less often	53.33%	46.67%

Table 5.11: Profile fit summary of survey path *Swipe method* and *Four-Items method*, presented for all participants and further segmented by gender, age group, and frequency of activities.

Overall Profile Fit

In the comparison between Swipe method and Four-Items method, 57.03% of participants preferred the profile generated by the Four-Items method, while 42.97% found the Swipe method's profile to be more representative. When comparing the Swipe method and Rating method, the Rating method was selected by a significantly larger number of

	Swipe	Rating
All	25.78%	74.22%
Gender		
Unknown	0%	100%
Female	29.33%	70.67%
Male	20.83%	79.17%
Non-binary	50%	50%
Age group		
Unknown	33.33%	66.67%
1-24	20.54%	79.46%
25-34	36.59%	63.41%
35-44	16.67%	83.33%
45-54	-	-
55-64	-	-
65+	-	-
Frequency of activities		
Unknown	9.09%	90.91%
Four times a week or more	30.30%	69.70%
Two to four times a month	25.00%	75.00%
Once a month or less often	22.22%	77.78%

Table 5.12: Profile fit summary of survey path *Swipe method and Rating method*, presented for all participants and further segmented by gender, age group, and frequency of activities.

participants (74.22%), whereas only 25.78% preferred the Swipe method's profile. In the third comparison between the Four-Items method and Two-Items method, the Two-Items method was slightly preferred, with 52.38% of participants selecting its generated profile, while 47.62% chose the profile from the Four-Items method.

To explore the differences in profile accuracy between elicitation methods, three pairwise Chi-square tests of independence were conducted. Bonferroni correction was applied to account for multiple comparisons, setting the adjusted significance threshold at $\alpha = 0.0167$. A significant difference was found between the Swipe method and the Four-Items method ($\chi^2(1) = 9.57, p = 0.002$), indicating that the elicitation method influenced profile selection. Similarly, a highly significant difference was observed between the Swipe method and the Rating method ($\chi^2(1) = 118.20, p < 0.001$). In contrast, no significant difference was found between the Four-Items method and the Two-Items method ($\chi^2(1) = 0.96, p = 0.327$), suggesting comparable outcomes between these two methods.

Overall, the findings highlight that while more expressive methods like the Rating method

	Four-items	Two-items
All	47.62%	52.38%
Gender		
Unknown	100%	0%
Female	46.67%	53.33%
Male	48.57%	51.43%
Non-binary	-	-
Age group		
Unknown	50%	50%
1-24	49.58%	50.42%
25-34	44.74%	55.26%
35-44	50%	50%
45-54	-	-
55-64	-	-
65+	-	-
Frequency of activities		
Unknown	25%	75%
Four times a week or more	53.85%	46.15%
Two to four times a month	47.57%	52.43%
Once a month or less often	46.67%	53.33%

Table 5.13: Profile fit summary of survey path *Four-Items method* and *Two-Items method*, presented for all participants and further segmented by gender, age group, and frequency of activities.

and Four-Items method can lead to higher perceived profile accuracy, simpler formats like Two-Items method may still be competitive.

Profile Fit by Gender

When analyzing profile fit by gender, among female participants, 50.67% selected the Four-Items method over the Swipe method, whereas 49.33% found the Swipe method's profile more representative. In the comparison between Swipe method and Rating method, 70.67% of female participants preferred the Rating method, while 29.33% selected the Swipe method. In the comparison between Four-Items method and Two-Items method, 53.33% of female participants preferred the Two-Items method over the Four-Items method (46.67%).

Male participants showed similar trends, with 68.89% selecting the Four-Items method over the Swipe method (31.11%) and 79.17% preferring the Rating method over the

Swipe method (20.83%). In the third comparison, 51.43% of male participants preferred the Two-Items method over the Four-Items method (48.57%).

Among non-binary participants, 100% selected the Swipe method over the Four-Items method, while in the comparison between Swipe method and Rating method, preferences were evenly split between the two methods. Participants who did not disclose their gender exclusively selected the Four-Items method (100%) over the Swipe method and the Two-Items method, while in survey path *Swipe method and Rating method* 100% preferred the Rating method in this group of participants. These findings suggest that gender-related differences in profile fit were relatively minor among female and male participants, who showed similar preferences across methods. The noticeable deviations observed among non-binary participants and those who did not disclose their gender are likely due to the small sample size.

Profile Fit by Age Group

Regarding age groups, participants aged 1-24 and 25-34 showed similar trends. Specifically, in both groups, the Four-Items method was preferred over the Swipe method, with 55.45% and 63.64% of participants selecting its profile compared to 44.55% and 36.36% selecting the profile of the Swipe method. In the comparison between Swipe method and Rating method, the Rating method was strongly preferred, with 79.46% of the 1-24 age group and 63.41% of the 25-34 age group selecting its profile, when 20.54% and 36.59%, respectively, have chosen the calculated profile of the Swipe method. In the comparison between Four-Items method and Two-Items method, 50.42% of the 1-24 group and 55.26% of the 25-34 group chose the Two-Items method's profile, while 49.58% of participants aged 1-24 and 44.74% aged 25-34 decided to choose the Four-Items method's profile.

The 35-44 age group exclusively selected the Swipe method over the Four-Items method while 83.33% of participants in this group preferred the Rating method over the Swipe method (16.67%). In the third comparison, the Four-Items method and Two-Items method were equally preferred in the age group 35-44.

No participants in the 45-64 age range provided responses. The single participant in the 65+ category preferred the Four-Items method over the Swipe method, but no data was available for the other comparisons. Participants who did not disclose their age group exclusively selected the Four-Items method (100%) over the Swipe method. 66.67% preferred the result of the Rating method over the Swipe method (33.33%). In the comparison between Four-Items method and Two-Items method, preferences were evenly split, with 50% selecting each method's profile.

These findings reveal consistent trends across the two largest age groups (1–24 and 25–34), both of which showed a clear preference for the Four-Items method and Rating method over the Swipe method. The Two-Items method and Four-Items method were nearly equally favored, indicating a balanced perception of profile quality. Due to the small number of responses in the other age groups and among participants who did not disclose their age, no definitive conclusions can be drawn for these segments.

Profile Fit by Frequency of Activities

Looking at profile fit by frequency of activities, participants who engage in leisure activities four times a week or more preferred the calculated profile of Four-Items method, with 57.14% of the participants over the profile of Swipe method (42.86%). 69.70% of this participant group preferred the Rating method's profile over the profile of the Swipe method (30.30%). In the comparison between the Four-Items method and Two-Items method, 53.85% of highly active participants chose the Four-Items method, while 46.15% preferred the Two-Items method.

Participants who engage in activities two to four times a month showed a preference for the Four-Items method (64.58%) over the Swipe method (35.42%), while strongly favoring the Rating method (75.00%) over the Swipe method (25.00%). In the comparison between Four-Items method and Two-Items method, 52.43% preferred the Two-Items method, compared to 47.57% of participants deciding for the Four-Items method.

Those who participate in leisure activities once a month or less often selected the Swipe method (53.33%) over the Four-Items method (46.67%), while in the comparison between Swipe method and Rating method, 77.78% found the Rating method's profile to be more representative and only 22.22% decided to choose the profile of the Swipe method. In the final comparison, 53.33% of these participants preferred the Two-Items method over the Four-Items method (46.67%).

Participants who did not disclose their frequency of activities showed a 75% preference for the Two-Items method over the Four-Items method, while in the Swipe method and Four-Items method comparison, 50% selected each profile. In the comparison between Swipe method and Rating method, they favored the Rating method (90.91%) over the Swipe method (9.09%).

These findings indicate that participants who are more actively engaged in leisure activities tended to prefer the profiles generated by more expressive methods such as the Four-Items method and Rating method. In contrast, those with lower activity frequency showed a slight preference for simpler methods like Swipe method and Two-Items method. This suggests that familiarity with leisure activities may influence the willingness to engage with more cognitively demanding elicitation types, potentially leading to more accurate and satisfying profile outcomes.

5.6 Effectiveness of Swipe-Based Onboarding on User Retention and Conversion

To evaluate the impact of a newly introduced interactive onboarding process on early user behavior, a cohort-based analysis was performed focusing on two key behavioral indicators: retention rate and conversion rate. The onboarding mechanism, introduced in calendar week 202329, was implemented through the Swipe method. Cohort identifiers in this analysis follow a standardized format, where the first four digits indicate the year of

data collection and the fifth and sixth digits correspond to the respective calendar week (e.g., 202329 = year 2023, week 29).

User cohorts from calendar weeks 202324 to 202328 served as the baseline group, representing system usage without onboarding. Starting from week 202329, the new onboarding feature was implemented. The analysis compared these two groups, focusing on the second week after initial use, a period often marked by early user drop-off.

5.6.1 Retention Rate

The onboarding group achieved in Week 2 an average retention rate of 13.80% (SD = 5.00%), whereas the baseline group without onboarding recorded a lower average of 8.40% (SD = 0.90%). A *t*-test resulted in a *t*-value of 2.34 and a *p*-value of 0.075. Although this result is not statistically significant ($p < 0.05$), it suggests a marginally significant positive trend in user retention attributable to the newly integrated onboarding process. This trend can also be observed in Figure A.9, which illustrates the retention distribution of the two groups in Week 2 as a boxplot. The onboarding group shows higher variability and a notably higher median retention (14.90%) compared to the lower median retention value of 8.00% in the group without onboarding. Further support for this observation is evident in the cohort retention heatmap (Figure A.10), where weekly retention is shown over time. While earlier cohorts without an onboarding process rarely exceeded 10.00%, cohorts in onboarding groups have reached retention rates of up to 20.00% in the same period.

Overall, these results indicate that, while not conclusively proven, the onboarding experience may have enhanced user engagement during the early stages of interaction.

5.6.2 Conversion Rate

Concurrently with the retention analysis, a cohort-based conversion analysis was conducted for the same period and cohort definitions. The results showed a significantly higher average conversion rate of 41.60% (SD = 24.50%) in the onboarding group compared to 5.10% (SD = 7.00%) in the baseline group. A *t*-test revealed a *t*-value of 3.20 with a *p*-value of 0.027, indicating a statistically significant difference between the two groups ($p < 0.05$). This provides strong evidence that the integration of the Swipe-based onboarding positively influenced early user conversion. The boxplot in Figure A.11 supports this result, with the onboarding group displaying both higher median conversion (33.33%) and greater variability compared to the group without onboarding showing a median conversion value of 0.00%. A more detailed view of weekly conversion rates across the cohorts is provided by the cohort-based heatmap in Figure A.12. This clearly indicates that cohorts in the onboarding group achieved substantially higher conversion rates reaching up to 77.00% compared to pre-onboarding cohorts, which reached a maximum of 14.00%. Furthermore, cohorts in the onboarding group demonstrated more sustained engagement over time.

These results highlight the effectiveness of the integrated Swipe-based preference elicitation method in the mobile app's onboarding process, demonstrating its ability to foster early user interaction and increase engagement.

CHAPTER 6

Discussion

The results of this study provide insights into the effectiveness of the four examined elicitation methods for addressing the CSP in a leisure activity RS. By comparing completion rates, time efficiency, usability and profile fit, this thesis highlights trade-offs between efficiency and accuracy in preference elicitation. The following chapter discusses these findings, compares them to previous research and outlines the limitations of this study.

6.1 Comparative Analysis of Elicitation Methods

Overall, the study results did not indicate a single distinctly preferable method. Instead, each method demonstrated specific strengths and weaknesses. The Swipe method, providing an intuitive interaction, resulted in the highest completion rate and highest usability rating among all methods. However, having the lowest profile fit indicates that binary decisions reduce the expressiveness needed to capture nuanced preferences. In comparison, the Rating method showed the highest profile accuracy and a competitive time efficiency. Nonetheless, its lower completion rate and reduced usability scores suggest that participants experienced higher cognitive load when using this method. The Two-Items method achieved a convincing balance between time efficiency and profile accuracy. However, it also exhibited the highest dropout rate, potentially due to the repetitive nature of binary decisions between similarly interesting options. Contrary to previous assumptions, the Four-Items method performed weakest overall. It was the slowest to complete, received the lowest usability ratings, and generated less accurate profiles than the Rating method and the Two-Items method.

6.2 Comparison With Related Work

The results of this study are consistent with previous findings in the field of preference elicitation in RSs. Research on binary input formats has emphasized, despite their intuitiveness and ease of use, they lack of informational depth and limit the expressiveness of captured preferences [SLL⁺13]. This aligns with the performance of the Swipe method, as it achieved the highest usability and completion rates, indicating that its low-effort and intuitive interactions are well-suited for mobile onboarding scenarios. At the same time, its low profile fit confirms the weakness of binary interactions in capturing the depth of user preferences. Furthermore, the high completion rate and perceived intuitiveness of the Swipe method can be explained through the Technology Acceptance Model (TAM). According to Davis [Dav89], perceived ease of use is a key indicator for the acceptance of new technologies. The simple and cognitively less demanding interaction design of the Swipe method reduces the barriers to engagement and increases the willingness of users to interact.

The high profile accuracy observed for the Rating method aligns with prior findings indicating that rating scales allow users to express more nuanced preferences [CGG⁺17]. At the same time, the study by Cena et al. [CGG⁺17] also highlights potential drawbacks, where such methods may introduce higher cognitive demands, leading to reduced engagement. This is also reflected in the results of the Rating method, where reduced engagement likely contributed to a lower completion rate.

The Two-Items method demonstrated strong time efficiency and accurate profile fit, which corresponds to earlier studies emphasizing the effectiveness of pairwise comparisons in capturing user preferences [GS10]. However, it also exhibited the lowest completion rate, potentially due to the repetitive nature of binary decisions between similarly interesting options.

In contrast, methods such as MaxDiff, which inspired the Four-Items method, have been shown to capture more detailed variations in user preferences [KM17]. Yet, the results of this study suggest that these theoretical advantages may not be fully transferable to mobile preference elicitation or to the leisure sector. The method's comparatively lower usability, slower completion time, and reduced profile fit indicate that visual complexity and increased cognitive effort can compensate the benefits of more expressive input formats.

The higher profile accuracies observed for the Two-Items method and Four-Items method can be explained through findings from Stated Choice research. Louviere et al. [LHS10] argue that relative decision tasks, as used in pairwise comparisons and MaxDiff approaches, lead to more stable and less biased preference expressions. By reducing cognitive uncertainty compared to absolute ratings, these methods contribute to the development of more robust user profiles, which is particularly essential during early onboarding phases of RSs.

These findings underline the importance of evaluating elicitation strategies not only in

terms of data quality, but also with regard to user experience and practical applicability. The analysis reveals an inherent trade-off between usability and profile accuracy. While simpler methods such as the Swipe method lower the entry barrier and lead to higher completion rates, more complex approaches like the Rating method or choice-based techniques enable a more differentiated and precise capture of individual preferences. In practical application, the choice of an elicitation method should be based on the platform's primary goals. If the focus is on quick onboarding and early user engagement, intuitive and low-effort methods are particularly advantageous. In contrast, if the long-term quality of recommendations and sustained user retention are of greater importance, more elaborate elicitation techniques can offer added value.

As the analysis of retention and conversion rates has shown, the integration of the Swipe-based onboarding method into the corporate partner's mobile app led to a measurable increase in early user engagement. This outcome highlights that in real-world deployments, simplicity in interaction design can be a decisive factor for success. It further emphasizes the need to align elicitation strategies with the technical and contextual constraints of the deployment environment and to carefully balance cognitive effort with the practical benefits of personalization.

6.3 Limitations

Some limitations must be considered that affect the significance and generalizability of the results. The study was conducted in the application context of a corporate partner in the leisure sector. This clearly defined context may limit the transferability of the results to other domains. The structure of the study design presents certain limitations, as well. While three different pairs of methods were compared, a comprehensive comparison of all possible combinations was not conducted. Therefore, potential interaction effects between specific methods may have remained unconsidered. Finally, it has to be mentioned that the survey was conducted in a browser-based prototype. This decision allowed for flexible participation, but may have influenced user behavior compared to native app usage.

CHAPTER 7

Conclusion

This final thesis chapter summarizes the main steps performed to obtain the key findings that answer the formulated research questions, highlights the scientific and practical contributions, and outlines directions for future research.

7.1 Summary

The aim of this study was to investigate the effectiveness of various visual preference elicitation methods in mitigating the CSP within a mobile RS for leisure activities. For this purpose, an extensive literature review was conducted, during which existing approaches to capturing user preferences were analyzed and compared. Based on this analysis, four commonly used methods were identified and selected for further comparison.

These four preference elicitation methods were implemented within a browser-based prototype that was used as part of a structured survey. Participants were randomly assigned, based on the least-used condition, to one of three survey paths and asked to evaluate leisure activities using the corresponding elicitation method. Each method was followed by a standardized questionnaire to assess the perceived usability. At the end of the session, participants were shown two preference profiles, which had been generated based on their previous inputs. They were then asked to select the profile that best reflected their actual interests.

The prototype was the result of an iterative development process that was systematically documented throughout this thesis. The data collected through the survey served as the foundation for answering RQ1, which focused on comparing the four methods.

The results of this evaluation were subsequently shared with the corporate partner. Based on these findings, the decision was made to integrate the Swipe-based method into the onboarding process of the mobile application. After a two-month observation period, the corporate partner provided anonymized usage data, which was used to answer RQ2

concerning the impact of visual preference elicitation on user retention and conversion in a real-world deployment.

7.2 Contributions

This section outlines the key contributions of the study in relation to the two research questions and highlights its relevance for both academic and practical contexts.

In response to RQ1, the results revealed that each method offers distinct strengths depending on the evaluated dimensions. The Swipe method achieved the highest completion rate and was perceived as the most user-friendly, making it particularly suitable for onboarding processes where simplicity and high participation rates are prioritized. The Rating method produced the most accurate preference profiles in terms of perceived profile fit, although it required more time and showed lower completion rates. The Two-Items method offered the highest time efficiency, balancing speed and profile accuracy, but suffered from a comparatively low completion rate. The Four-Items method, while conceptually promising, did not outperform the other methods in any of the measured dimensions. Due to its longer completion time and higher cognitive demand, it may not be suitable for mobile or quick onboarding processes in its implemented form.

In response to RQ2, the comparative analysis with baseline groups without visual onboarding showed that the Swipe-based visual preference elicitation implemented by the corporate partner positively influenced both retention and conversion rates. Compared to the baseline groups without visual onboarding, the onboarding groups exhibited a higher median retention rate (14.90% compared to 8.00%) and a clearly higher median conversion rate (33.30% compared to 0.00%).

From a practical perspective, these findings provide guidance for mobile application designers in selecting preference elicitation methods that align with their platform's goals, whether the priority is time efficiency, engagement, or recommendation precision. Different methods could be implemented and tested against each other in A/B tests to evaluate their effectiveness under real-world conditions and to identify the most suitable approach for the given context.

While the results indicate clear trends, the findings must be interpreted in light of the limitations discussed in Chapter 6.3. In addition to those already outlined, further factors should be taken into account. These include the restricted observation period of two months and the use of non-random snowball sampling. Together with the limitations already mentioned, such as the domain-specific application context, the limited comparison of method combinations, and the use of a browser-based prototype, these aspects may influence the generalizability of the results and should be considered when transferring the findings to other domains.

This study contributes to both academic research and practical development in the domain of RSs and user onboarding by offering a comparative analysis of four visual preference elicitation methods. These were implemented in a functional browser-based prototype

designed for a real-world use case that was specifically developed for this thesis project. This analysis provides valuable insights into the strengths and weaknesses of each method in a practical setting. In addition the study delivers empirical insights into how usability and cognitive load influence user behavior during the onboarding process, highlighting the importance of balancing these factors to enhance user engagement and satisfaction. For practitioners, the thesis delivers valuable insights into designing onboarding flows that effectively balance user effort and recommendation accuracy, providing strategies to enhance user satisfaction and engagement in real-world applications. Moreover, the findings underline that the choice of a preference elicitation method affects not only initial user loyalty, but also the quality of the resulting recommendations, which in turn has a significant influence on long-term user retention.

7.3 Future Work

To further develop the findings of this study, future research could aim to compare all examined elicitation methods directly against each other within a unified experimental design. This would allow for a more robust and statistically conclusive assessment of their relative strengths and weaknesses. In addition, a more detailed evaluation of user experience could be achieved by employing an extended usability questionnaire that includes established constructs such as choice satisfaction, next to the examined constructs choice difficulty and ease of use. This would enable a more nuanced understanding of how users perceive not only the usability of an elicitation method, but also the quality and clarity of their decision-making processes.

Besides the quantitative results, the qualitative comments provided by participants offered valuable insights into suggestions for improving the prototype. A frequent point of criticism concerned the generality of the activity categories, which were often perceived as too broad or unclear. One participant stated: *"[Ich konnte mit nur einer der Aktivitäten etwas anfangen, die anderen Begriffe waren sehr allgemein gehalten. Es war meist eine Wahl zwischen Aktivitäten, die ich nicht mache und die ich nicht machen werde.] – I could only relate to one of the activities; the other terms were too general. Most of the time, it was a choice between activities that I do not engage in and will not engage in."*

Furthermore, the answer options relating to the demographic dimension *frequency of leisure activities* were considered insufficient by several participants. As one comment indicated: *"[Die Antwortmöglichkeiten bei der Frage, wie oft unternimmst du etwas, waren mau. Etwas zwischen 4x/Woche und 1-4x/Monat wäre noch sinnvoll gewesen.] – The answer options for the question about how often you engage in activities were poor. Something between 4x/week and 1-4x/month would have made sense."*

Another important aspect highlighted was the lack of technical functionalities, particularly the inability to correct inputs once submitted. One participant emphasized: *"[Ein Zurück Button ist unbedingt notwendig. Ich habe Angaben gemacht, die nicht richtig waren. Aber einmal geswiped und Auswahl konnte ich nicht rückgängig machen.] – A back button is*

absolutely necessary. I made selections that were incorrect, but once swiped, I could not undo the choice."

Beyond methodological refinements, the observation period could be extended beyond the two-month timeframe used in this study. A longer observation window would provide deeper insights into how onboarding strategies influence long-term user engagement and retention. This would allow researchers to examine whether the quality of initial preference elicitation has lasting effects on user retention and engagement over time.

Since this study was conducted in the context of leisure activity recommendations, it would also be of interest to investigate how the evaluated elicitation methods perform in other domains. A cross-domain application could help determine whether the results of this study are domain-specific or reflect more generalizable patterns in user interaction and onboarding success.

Another interesting perspective for future research would be the development of adaptive elicitation strategies that dynamically combine different methods. For example, an initial preference elicitation could be performed using a Swipe-based interface to ensure a low interaction barrier. After the initial preference elicitation, the system could dynamically move to more precise methods such as pairwise comparisons to gradually improve profile accuracy. Furthermore, it seems promising to develop adaptive systems that are able to switch between different elicitation methods in real time depending on user behavior, interaction duration or dropout probability. While some research exists on adaptive preference elicitation and hybrid methods, the specific combination of visual elicitation methods may be of particular interest in the context of RSs and mobile devices.

Overall, this study shows that the examined visual preference elicitation methods vary in their strengths depending on the intended onboarding experience and the identification of accurate preference profiles. As revealed through the integration of the Swipe-based method by the corporate partner, simplicity in interaction design can be a decisive factor for success. Therefore, method selection should not follow a one-size-fits-all approach, but should instead be carefully aligned with the specific goals and contextual conditions of the application.

APPENDIX A

Figures

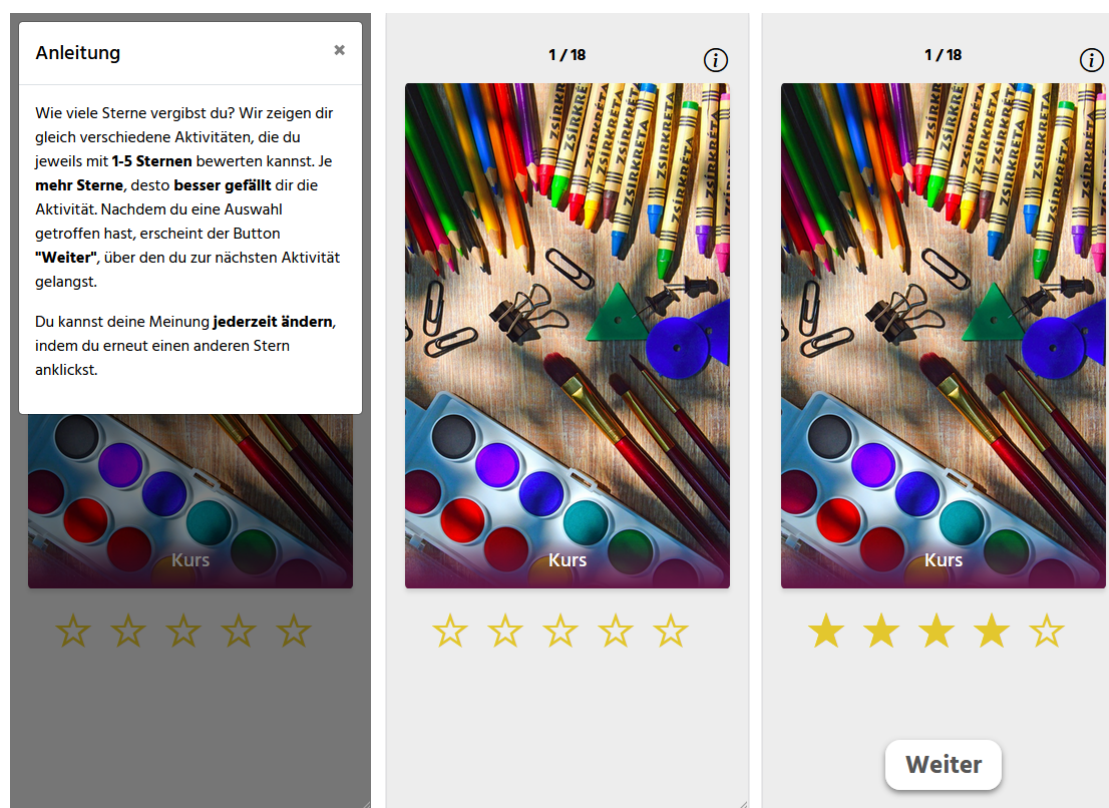


Figure A.1: Visual representation of Rating method. On the left side when entering this method with instructional pop-up, in the middle in initial state and on the right side when a rating has been chosen.

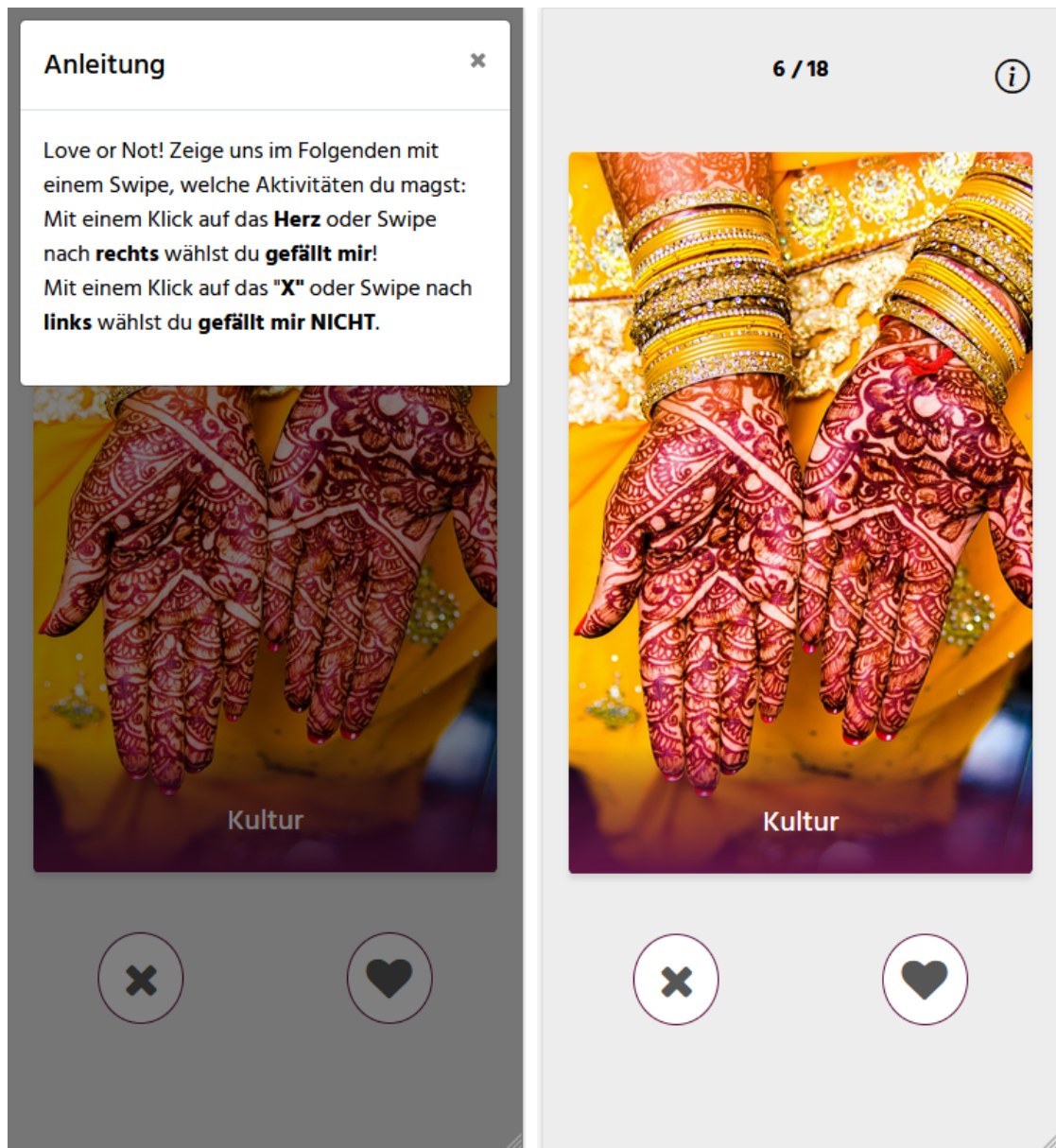


Figure A.2: Visual representation of Swipe method. On the left side when entering this method with instructional pop-up and on the right side in initial state.

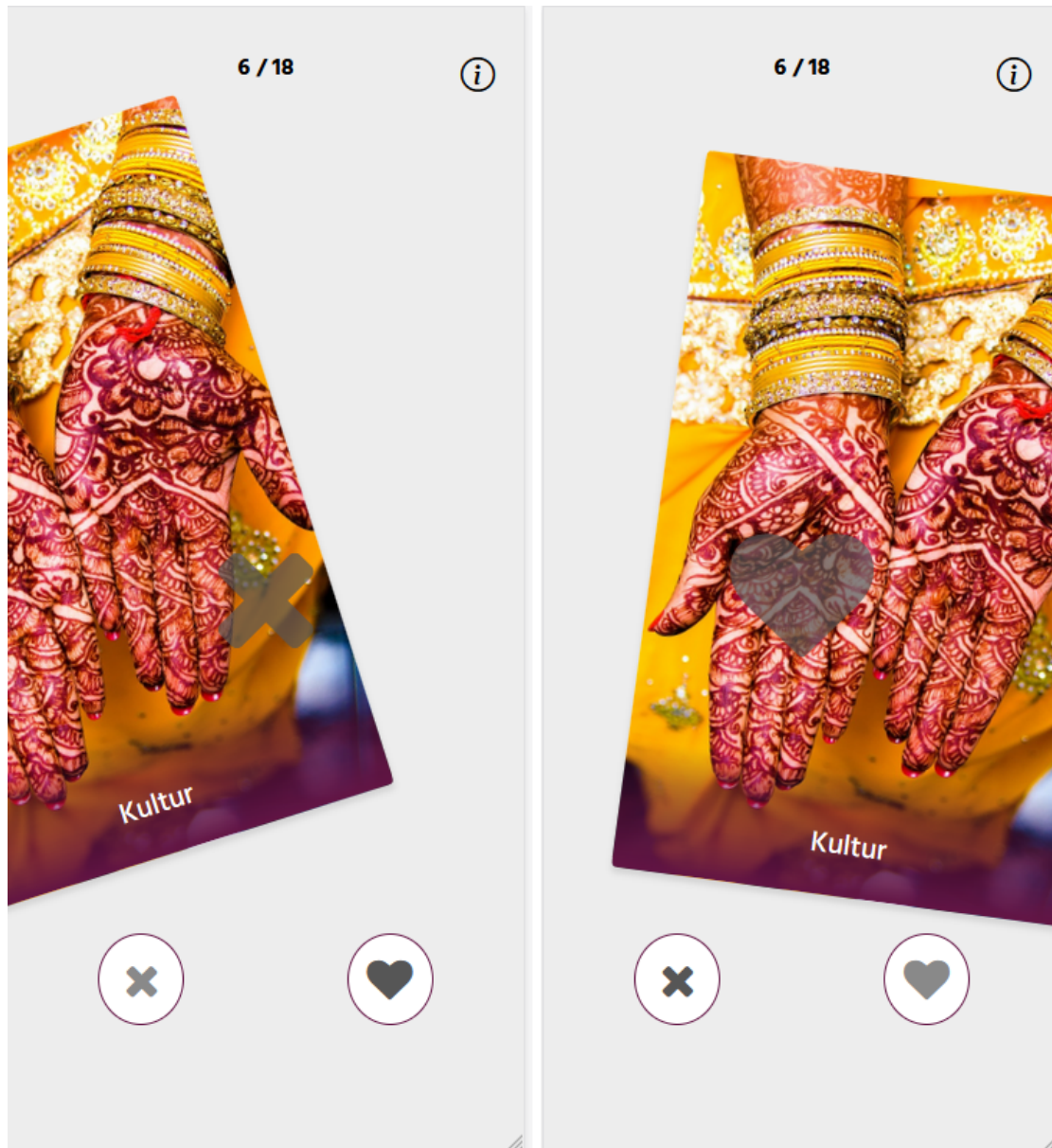


Figure A.3: Visual representation of Swipe method. On the left side if a negative rating is given by swiping and on the right side if a positive rating is given by swiping.

A. FIGURES

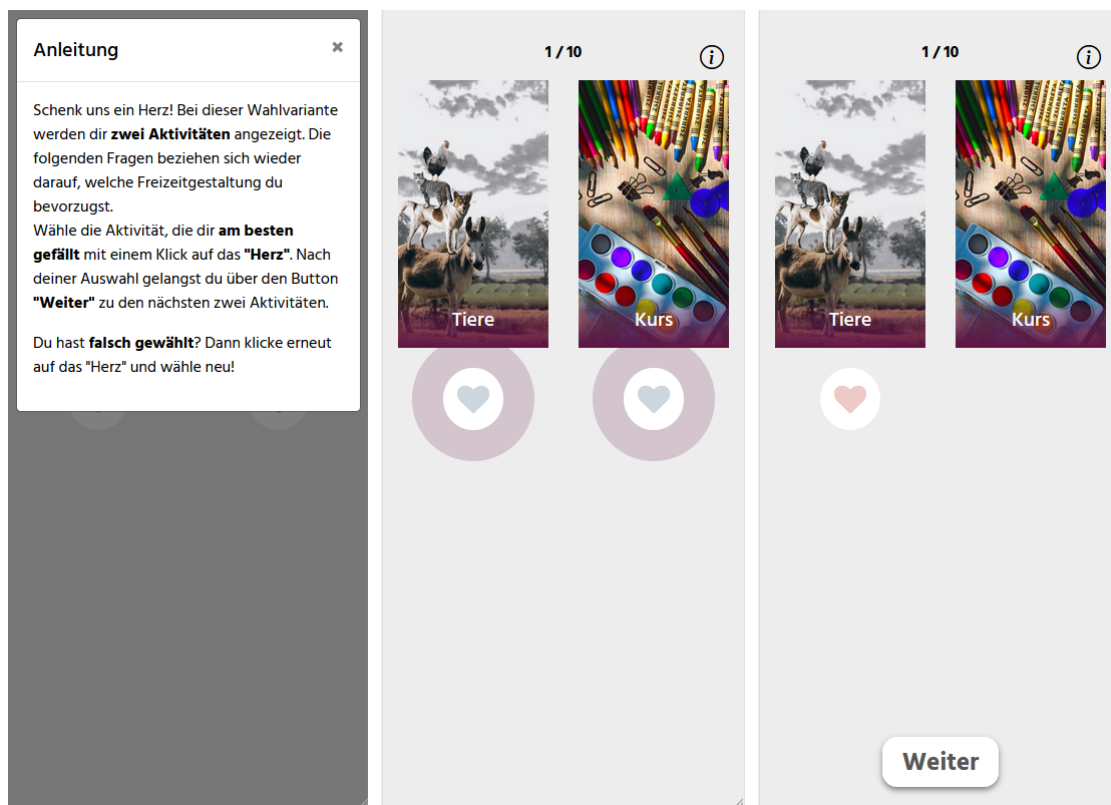


Figure A.4: Visual representation of Two-Items method. On the left side when entering this method with instructional pop-up, in the middle in initial state with pulsating interaction elements and on the right side when a rating has been chosen.

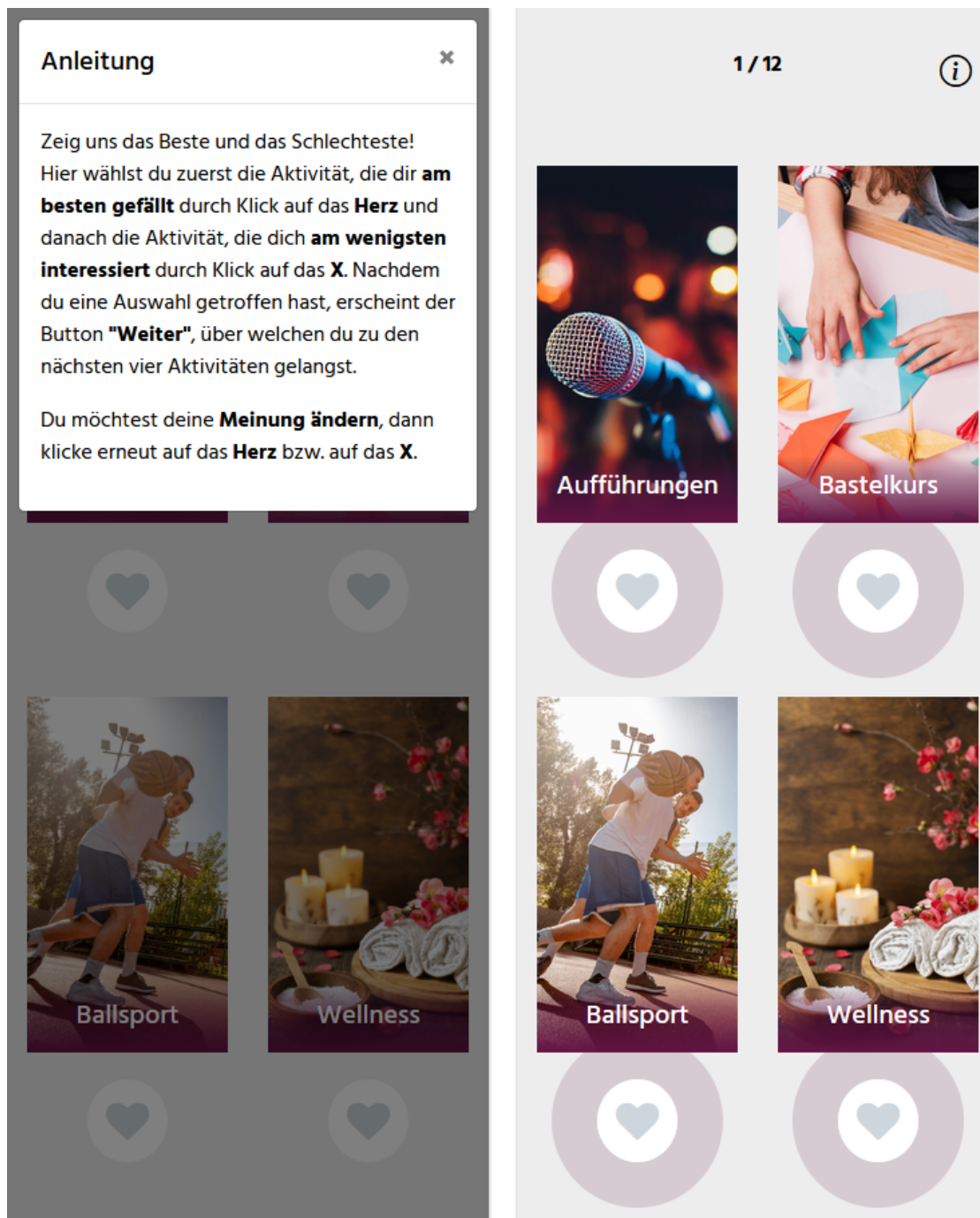


Figure A.5: Visual representation of Four-Items method. On the left side when entering this method with instructional pop-up and on the right side in initial state with pulsating interaction elements.

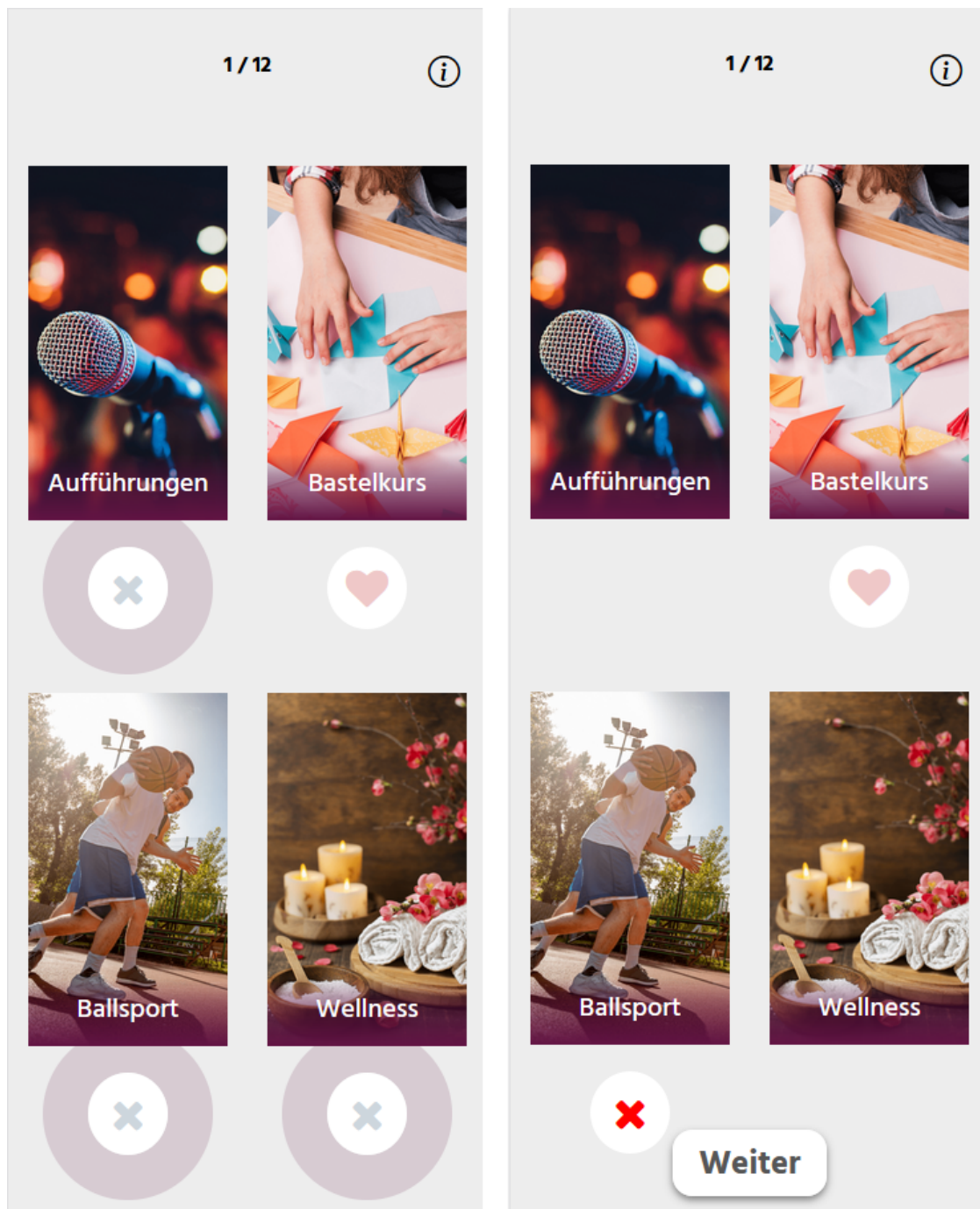


Figure A.6: Visual representation of Four-Items method. On the left side when positive rating has been chosen and on the right side when both positive and negative rating has been chosen.

A. Die folgenden Fragen beziehen sich auf die **Teilfragen**, die du entweder über den **Swipe**, oder durch das Klicken auf das **Herz** bzw. **X** ausgefüllt hast. Bitte gib an, wie sehr du den folgenden Aussagen zustimmst.

Es fiel mir leicht, mich für die jeweiligen Antwortmöglichkeiten zu entscheiden.

☐ stimme überhaupt nicht zu ☐ stimme nicht zu ☐ stimme weder zu noch nicht zu ☐ stimme zu ☐ stimme voll zu

Sich für die jeweiligen Antwortmöglichkeiten zu entscheiden, hat nicht lange gedauert.

☐ stimme überhaupt nicht zu ☐ stimme nicht zu ☐ stimme weder zu noch nicht zu ☐ stimme zu ☐ stimme voll zu

B. Die folgenden Fragen beziehen sich auf den **gesamten Prozess**, also wie du alle Schritte wahrgenommen hast. Bitte gib an, wie sehr du den folgenden Aussagen zustimmst.

Ich empfand die gesamte Herangehensweise als sehr benutzerfreundlich.

☐ stimme überhaupt nicht zu ☐ stimme nicht zu ☐ stimme weder zu noch nicht zu ☐ stimme zu ☐ stimme voll zu

Die Nutzung war leicht verständlich.

☐ stimme überhaupt nicht zu ☐ stimme nicht zu ☐ stimme weder zu noch nicht zu ☐ stimme zu ☐ stimme voll zu

Die Durchführung war langwierig.

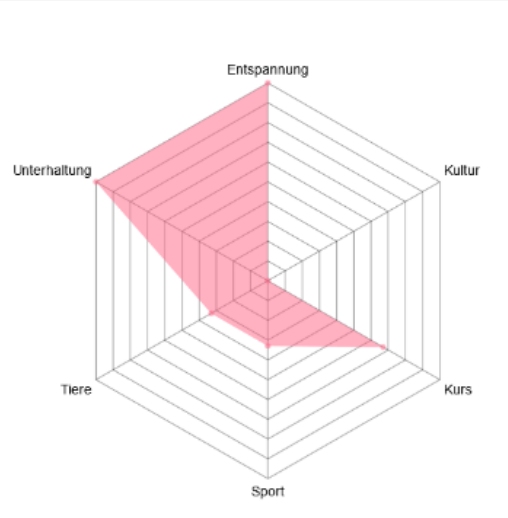
☐ stimme überhaupt nicht zu ☐ stimme nicht zu ☐ stimme weder zu noch nicht zu ☐ stimme zu ☐ stimme voll zu

Weiter

Figure A.7: Visual representation of questionnaire with propositions for determining perceived usability.

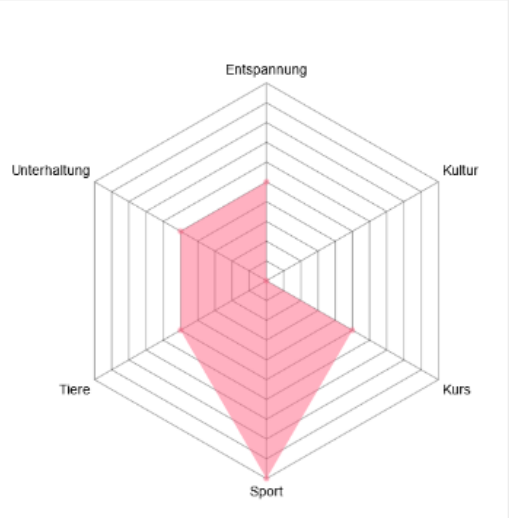
A. FIGURES

**Wir haben zwei möglich Profile für dich.
Welches passt besser zu dir?**



Platz 1	Entspannung, Unterhaltung
Platz 2	Kurs
Platz 3	Sport, Tiere
Platz 4	Kultur

☐ Dieses Profil entspricht eher meinen Präferenzen



Platz 1	Sport
Platz 2	Entspannung, Kurs, Tiere, Unterhaltung
Platz 3	Kultur

☐ Dieses Profil entspricht eher meinen Präferenzen

Wie alt bist du?

☒ Keine Angabe
 ☐ 1-24 Jahre
 ☐ 25-34 Jahre
 ☐ 35-44 Jahre
 ☐ 45-54 Jahre
 ☐ 55-64 Jahre
 ☐ 65+ Jahre

Was ist dein Geschlecht?

☒ Keine Angabe
 ☐ Männlich
 ☐ Weiblich
 ☐ Intersexuell/Divers

Wie oft unternimmst du was?

☒ Keine Angabe
 ☐ viermal pro Woche oder öfter
 ☐ zwei- bis viermal im Monat
 ☐ einmal im Monat oder seltener

Du möchtest an unserem Gewinnspiel mitmachen?

Dann klicke auf Ja und akzeptiere die unten angegebenen Teilnahmebedingungen. Außerdem brauchen wir deine E-Mailadresse, um dich zu kontaktieren, solltest du gewonnen haben.

☐ Ja, ich möchte beim Gewinnspiel teilnehmen und akzeptiere die Teilnahmebedingungen.

Absenden

Figure A.8: Visual representation of the most relevant preference profile selection and (optional) demographic information.

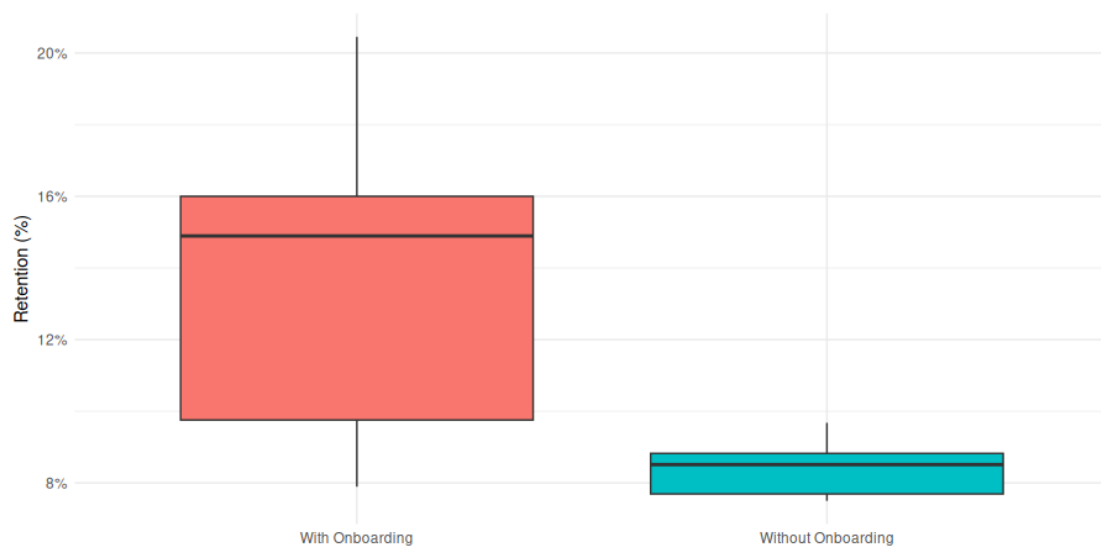


Figure A.9: Comparison of user retention with and without onboarding in Week 2.

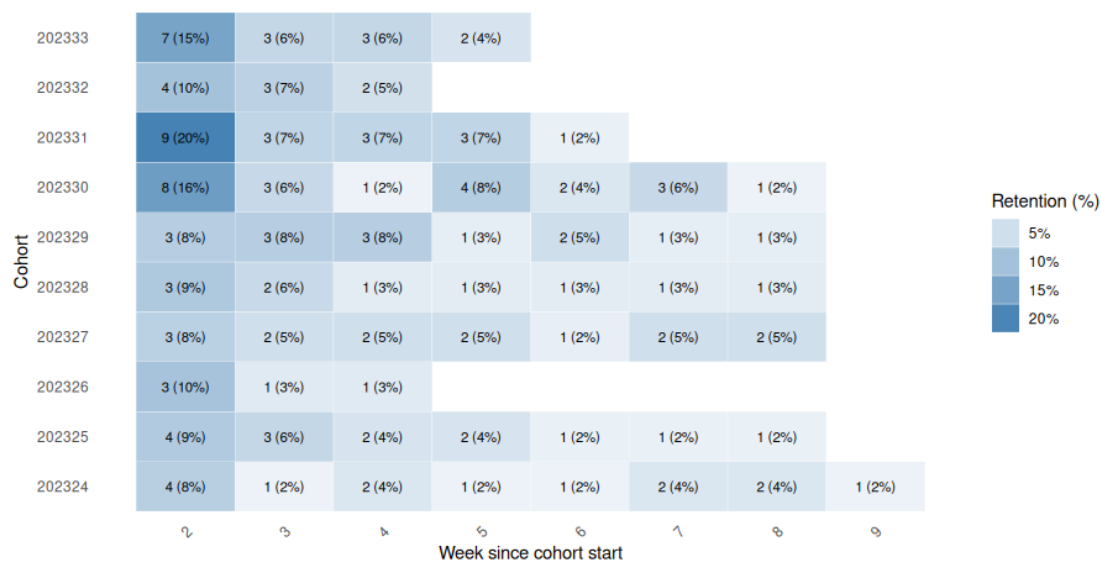


Figure A.10: Cohort analysis of user retention in observed period.

A. FIGURES

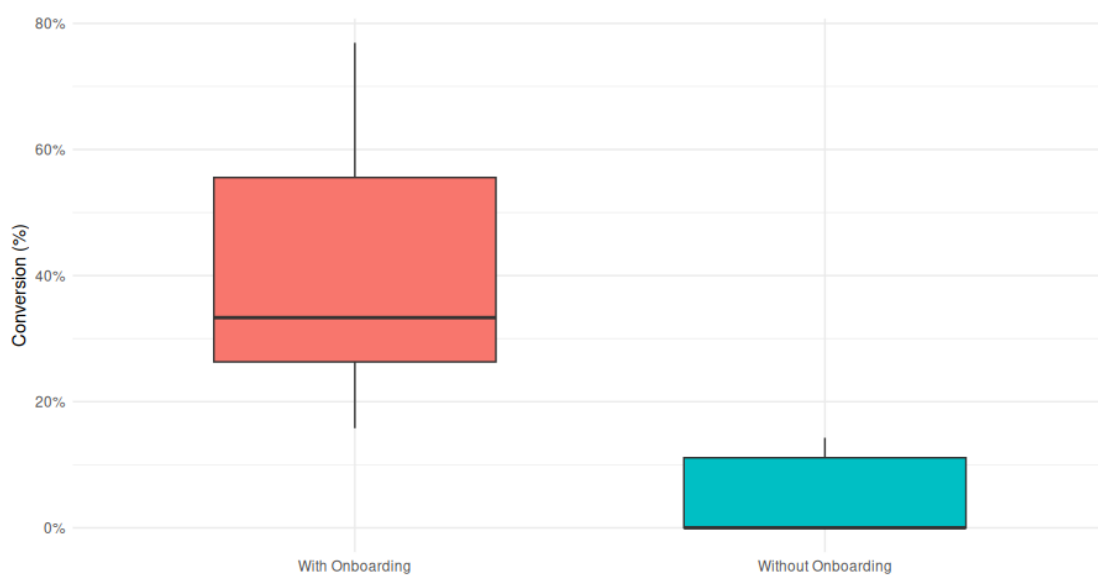


Figure A.11: Comparison of user conversion with and without onboarding in Week 2.

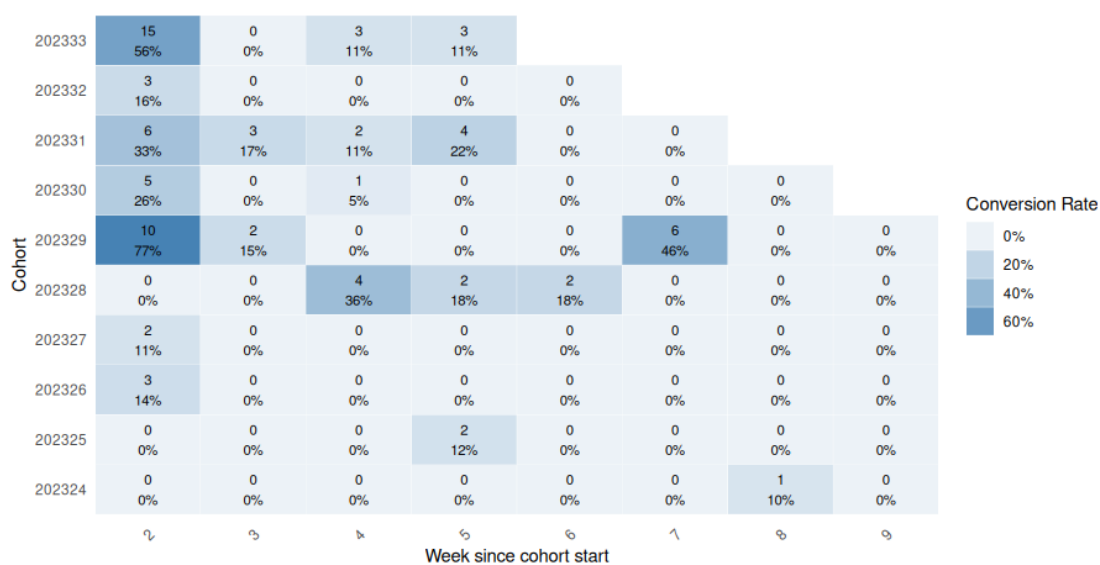


Figure A.12: Cohort analysis of user conversion in observed period.

Overview of Generative AI Tools Used

In this thesis, ChatGPT by OpenAI was used as a supporting tool. In addition to the search engine Google, it was used to answer general questions. This included, for example, questions regarding the functionality of R and LaTeX to gain a quicker understanding than it would have been possible by reading the respective documentation. During the writing process, ChatGPT was used to check the grammar of self-written English sentences and to rephrase them when necessary to improve readability. Furthermore, ChatGPT was used to generate code snippets for statistical analysis in R. All outputs generated by ChatGPT were reviewed, checked for validity, and corrected if necessary, to the best of my knowledge.

List of Figures

3.1	Visual representation of the survey flow.	16
5.1	Duration distributions of completed elicitation methods.	36
5.2	Duration distributions of completed elicitation methods of different genders.	38
5.3	Duration distributions of completed elicitation methods of different age groups.	39
5.4	Duration distributions of completed elicitation methods of different activity frequencies.	39
A.1	Visual representation of Rating method. On the left side when entering this method with instructional pop-up, in the middle in initial state and on the right side when a rating has been chosen.	59
A.2	Visual representation of Swipe method. On the left side when entering this method with instructional pop-up and on the right side in initial state.	60
A.3	Visual representation of Swipe method. On the left side if a negative rating is given by swiping and on the right side if a positive rating is given by swiping.	61
A.4	Visual representation of Two-Items method. On the left side when entering this method with instructional pop-up, in the middle in initial state with pulsating interaction elements and on the right side when a rating has been chosen.	62
A.5	Visual representation of Four-Items method. On the left side when entering this method with instructional pop-up and on the right side in initial state with pulsating interaction elements.	63
A.6	Visual representation of Four-Items method. On the left side when positive rating has been chosen and on the right side when both positive and negative rating has been chosen.	64
A.7	Visual representation of questionnaire with propositions for determining perceived usability.	65
A.8	Visual representation of the most relevant preference profile selection and (optional) demographic information.	66
A.9	Comparison of user retention with and without onboarding in Week 2.	67
A.10	Cohort analysis of user retention in observed period.	67
A.11	Comparison of user conversion with and without onboarding in Week 2.	68
A.12	Cohort analysis of user conversion in observed period.	68

List of Tables

5.1	Participants grouped by device types.	33
5.2	Number of participants who completed respectiv survey path.	34
5.3	Participants grouped by age ranges.	34
5.4	Participants grouped by gender.	35
5.5	Participants grouped by frequency of activities.	35
5.6	Completion rate in percent per elicitation method.	36
5.7	Average duration (in seconds) summary across four elicitation methods, presented for all participants and further segmented by gender, age group, and frequency of activities.	37
5.8	Results of the Dunn test with Bonferroni correction for pairwise comparisons of the methods. Significant differences with $p < 0.05$	40
5.9	Average usability (1 to 5 Likert-Scale) summary across four elicitation methods, presented for all participants and further segmented by gender, age group, and frequency of activities.	41
5.10	Results of the Dunn test with Bonferroni correction for pairwise comparisons of the methods. Significant differences with $p < 0.05$	42
5.11	Profile fit summary of survey path <i>Swipe method and Four-Items method</i> , presented for all participants and further segmented by gender, age group, and frequency of activities.	43
5.12	Profile fit summary of survey path <i>Swipe method and Rating method</i> , presented for all participants and further segmented by gender, age group, and frequency of activities.	44
5.13	Profile fit summary of survey path <i>Four-Items method and Two-Items method</i> , presented for all participants and further segmented by gender, age group, and frequency of activities.	45

Glossary

Four-Items method A choice-based elicitation method in which users are presented four leisure activities and asked to select the one they like most and the one they like least. 14, 15, 18, 20, 23, 24, 26, 28, 30–32, 34–47, 51, 52, 56, 63, 64, 71, 73

Rating method An ordinal elicitation method in which users indicate their preference for a displayed leisure activity by selecting a value on a five-point scale, represented by one to five stars. 14, 15, 18, 20, 23, 27, 34–47, 51–53, 56, 59, 71, 73

Swipe method A binary elicitation method in which users indicate their preference for a displayed leisure activity by swiping left or right or by selecting predefined symbols representing "like" or "dislike". 14, 15, 18, 19, 23, 27, 31, 34–38, 40–47, 51–53, 56, 60, 61, 71, 73

tuple Refers to a predefined set of four distinct leisure activity items used in the Four-Items Method, generated to ensure balanced and diverse presentation of item combinations during the elicitation proces.. 30, 31

Two-Items method A choice-based elicitation method in which users are presented two leisure activities and select the one they prefer. 14, 15, 18, 20, 23, 24, 27–29, 32, 34–37, 40–42, 44–47, 51, 52, 56, 62, 71, 73

xpo-elicitation-item-answer A database table that stores user responses for each evaluated item, including the item ID, the assigned rating, the timestamp, and the associated survey path. 23

xpo-elicitation-mode A database table that stores the predefined survey paths and tracks the number of times each path has been assigned to ensure balanced distribution among participants. 15

xpo-elicitation-question-answer A database table that stores user responses to questionnaire propositions, including the user ID, the specific proposition, the selected rating, the timestamp, and the survey step in which the questionnaire was completed. 24

xpo-elicitation-user A database table that stores individual participant records, including their assigned elicitation path and demographic information, if disclosed.
15, 16

Acronyms

- CBF** content-based filtering. 1, 8, 9
- CF** collaborative filtering. 1, 8, 10
- CI** Corporate Identity. 26
- CNN** Convolutional neural network. 9, 12
- CRS** Conversational Recommender System. 11
- CSP** cold start problem. 1–3, 6–10, 13, 51, 55
- DSR** Design Science Research. 4, 13, 31
- GUI** Graphical User Interface. 25
- HCI** Human-Computer Interaction. 26
- ICS** item cold start. 9
- MaxDiff** Maximum Difference Scaling. 14, 52
- MF** matrix factorization. 9, 10
- NCF** Neural collaborative filtering. 10
- POI** points of interest. 12
- RNN** Recurrent neural network. 9
- RQ1** Research Question 1. 2, 17, 55, 56
- RQ2** Research Question 2. 3, 55, 56
- RS** recommender system. 1–4, 6–11, 13, 17, 18, 21, 25, 51, 52, 55, 56, 58

SCS system cold start. 9

SVD Singular Value Decomposition. 9

UCS user cold start. 9

Bibliography

- [AG21] Fadi AlMahamid and Katarina Grolinger. Reinforcement learning algorithms: An overview and classification. In *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–7, 2021.
- [Agg16] Charu C. Aggarwal. *Recommender Systems*. Springer International Publishing, 2016.
- [AT05] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17:734–749, 6 2005.
- [BA09] Linas Baltrunas and Xavier Amatriain. Towards time-dependant recommendation based on implicit feedback. In *Workshop on context-aware recommender systems (CARS’09)*, pages 25–30, 2009.
- [BB23] Fjolla Berisha and Eliot Bytyçi. Addressing cold start in recommender systems with neural networks: a literature survey. *International Journal of Computers and Applications*, 45:485–496, 8 2023.
- [Bur00] Robin Burke. Knowledge-based recommender systems. *Encyclopedia of library and information systems*, 69:175–186, 2000.
- [Bur07] Robin Burke. *Hybrid Web Recommender Systems*, pages 377–408. Springer, 2007.
- [CCF⁺06] Sara J Czaja, Neil Charness, Arthur D Fisk, Christopher Hertzog, Sankaran N Nair, Wendy A Rogers, and Joseph Sharit. Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (create). Technical report, 2006.
- [CFY⁺21] Andrés Cimadamore, Alejandro Fernandez, Chenhui Ye, Pascale Zaraté, and Daouda Kamissoko. *A User Interface for Consistent AHP Pairwise Comparisons*, pages 119–134. 2021.

- [CGG⁺17] Federica Cena, Cristina Gena, Pierluigi Grillo, Tsvi Kuflik, Fabiana Vernerio, and Alan J. Wecker. How scales influence user rating behaviour in recommender systems. *Behaviour and Information Technology*, 36:985–1004, 10 2017.
- [CP09] Li Chen and Pearl Pu. Interaction design guidelines on critiquing-based recommender systems. *User Modeling and User-Adapted Interaction*, 19:167–206, 8 2009.
- [CPB21] Alina Cosma, Jan Pavelka, and Petr Badura. Leisure time use and adolescent mental well-being: Insights from the covid-19 czech spring lockdown. *International Journal of Environmental Research and Public Health*, 18, 12 2021.
- [CT11] Li Chen and Ho Keung Tsoi. Users’ decision behavior in recommender interfaces: Impact of layout design. In *RecSys’ 11 Workshop on Human Decision Making in Recommender Systems*, 2011.
- [Dav89] Fred D. Davis. Technology acceptance model: Tam. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption*, 205:5, 1989.
- [FHH00] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. Measuring usability. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 345–352. ACM, 4 2000.
- [GJ17] Jyotirmoy Gope and Sanjay Kumar Jain. A survey on solving cold start problem in recommender systems. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 133–138. IEEE, 5 2017.
- [GS10] Shengbo Guo and Scott Sanner. Real-time multiattribute bayesian preference elicitation with pairwise comparison queries. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 289–296. JMLR Workshop and Conference Proceedings, 2010.
- [GT11] Adomavicius Gediminas and Alexander Tuzhilin. *Context-Aware Recommender Systems*, pages 217–253. Springer US, 2011.
- [GV17] Satya Keerthi Gorripati and Valli Kumari Vatsavayi. Community-based collaborative filtering to alleviate the cold-start and sparsity problems. Technical report, 2017.
- [GW15] Mark P. Graus and Martijn C. Willemsen. Improving the user experience during cold start through choice-based preference elicitation. In *RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems*, pages 273–276. Association for Computing Machinery, Inc, 9 2015.

- [Hev07] Alan R Hevner. A three cycle view of design science research. Technical report, 2007.
- [HSvEB21] Sebastian Himmler, Vikas Soekhai, Job van Exel, and Werner Brouwer. What works better for preference elicitation among older people? cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting. *Journal of Choice Modelling*, 38, 3 2021.
- [ILN⁺21] Andrea Iovine, Pasquale Lops, Fedelucio Narducci, Marco de Gemmis, and Giovanni Semeraro. Improving preference elicitation in a conversational recommender system with active learning strategies. In *Proceedings of the 36th Annual ACM symposium on applied computing*, pages 1375–1382, 2021.
- [JBB11] Nicolas Jones, Armelle Brun, and Anne Boyer. Comparisons instead of ratings: Towards more stable preferences. In *Proceedings - 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011*, volume 1, pages 451–456, 2011.
- [JGJ⁺22] Syed Irteza Hussain Jafri, Rozaida Ghazali, Irfan Javid, Zahid Mahmood, and Abdullahi Abdi Abubakar Hassan. Deep transfer learning with multimodal embedding to tackle cold-start and sparsity issues in recommendation system. *PLOS ONE*, 17:e0273486, 8 2022.
- [JN11] Kevin G Jamieson and Robert D Nowak. Active ranking using pairwise comparisons. In *Advances in neural information processing systems 24*, 2011.
- [JZ12] Dietmar Jannach and Markus Zanker. *Value and Impact of Recommender Systems*, pages 519–546. Springer US, 2012.
- [Kim21] Kihwan Kim. An empirical analysis on transparent algorithmic exploration in recommender systems. *arXiv preprint arXiv:2108.00151*, 2021.
- [KM16] Svetlana Kiritchenko and Saif M Mohammad. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2016.
- [KM17] Svetlana Kiritchenko and Saif M Mohammad. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2017), Vancouver, Canada, 2017*, 2017.
- [KPG22] Nikolai Krivulin, Alexey Prinkov, and Igor Gladkikh. Using pairwise comparisons to determine consumer preferences in hotel selection. *Mathematics*, 10:730, 2 2022.

- [KWG⁺12] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22:441–504, 10 2012.
- [KWH10] Bart P Knijnenburg, Martijn C Willemsen, and Stefan Hirtbach. Receiving recommendations and providing feedback: The user-experience of a recommender system. In *E-Commerce and Web Technologies: 11th International Conference, EC-Web 2010, Bilbao, Spain, September 1-3, 2010. Proceedings 11*, pages 207–216, 2010.
- [KZFM09] Evangelos Karapanos, John Zimmerman, Jodi Forlizzi, and Jean-Bernard Martens. User experience over time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 729–738. ACM, 4 2009.
- [LHS10] Jordan J.. Louviere, David A.. Hensher, and Joffre Dan. Swait. *Stated choice methods : analysis and applications*. Cambridge University Press, 2010.
- [LHZ14] Benedikt Loepp, Tim Hussein, and Jürgen Ziegler. Choice-based preference elicitation for collaborative filtering recommender systems. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 3085–3094. Association for Computing Machinery, 2014.
- [LHZZ24] Mingming Li, Songlin Hu, Fuqing Zhu, and Qiannan Zhu. Few-shot learning for cold-start recommendation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7185–7195, 2024.
- [LKS⁺14] Emanuel Lacic, Dominik Kowald, Paul Seitlinger, Christoph Trattner, and Denis Parra. Recommending items in social tagging systems using tag and time information. 6 2014.
- [MGW21] Yutao Ma, Xiao Geng, and Jian Wang. A deep neural network with multiplex interactions for cold-start service recommendation. *IEEE Transactions on Engineering Management*, 68:105–119, 2 2021.
- [MID23] Loubna Mekouar, Youssef Iraqi, and Issam Damaj. A global user profile framework for effective recommender systems. *Multimedia Tools and Applications*, 83:50711–50731, 5 2023.
- [MP09] Jeremy Mendel and Richard Pak. The effect of interface consistency and cognitive load on user performance in an information search task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 53:1684–1688, 10 2009.
- [MSdGL16] Cataldo Musto, Giovanni Semeraro, Marco de Gemmis, and Pasquale Lops. *Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems*, pages 729–734. 2016.

- [NSSW15] Julia Neidhardt, Leonhard Seyfang, Rainer Schuster, and Hannes Werthner. A picture-based approach to recommender systems. *Information Technology and Tourism*, 15:49–69, 3 2015.
- [Ovi06] Sharon Oviatt. Human-centered design meets cognitive load theory. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 871–880. ACM, 10 2006.
- [PBW⁺12] Alina Pommeranz, Joost Broekens, Pascal Wiggers, Willem Paul Brinkman, and Catholijn M. Jonker. Designing interfaces for explicit preference elicitation: A user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction*, 22:357–397, 10 2012.
- [PC08] Pearl Pu and Li Chen. User-involved preference elicitation for product search and recommender systems. *AI magazine*, pages 93–103, 2008.
- [RD22] Deepjyoti Roy and Mala Dutta. A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9, 12 2022.
- [RGR24] Amir Mohammad Rahmani, Wim Groot, and Hamed Rahmani. Dropout in online higher education: a systematic literature review, 12 2024.
- [RRS10] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to Recommender Systems Handbook*, pages 1–35. Springer US, 2010.
- [SL16] Jeff Sauro and James R. Lewis. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.
- [SLL⁺13] Mingxuan Sun, Fuxin Li, Joonseok Lee, Ke Zhou, Guy Lebanon, and Hongyuan Zha. Learning multiple-question decision trees for cold-start recommendation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 445–454. ACM, 2 2013.
- [SNW20] Mete Sertkan, Julia Neidhardt, and Hannes Werthner. Eliciting touristic profiles: A user study on picture collections. In *UMAP 2020 - Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 230–238. Association for Computing Machinery, Inc, 7 2020.
- [SS17] Kalpathy Ramaiyer Subramanian and K R Subramanian. Product promotion in an era of shrinking attention span. *International Journal of Engineering and Management Research*, pages 85–91, 2017.
- [TV11] Stamatina Thomaidou and Michalis Vazirgiannis. Multiword keyword recommendation system for online advertising. In *Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011*, pages 423–427, 2011.

- [VM00] Viswanath Venkatesh and Michael G. Morris. Why don't men ever stop to ask for directions? gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly*, 24:115, 3 2000.
- [WL22] Hongwei Wang and Jure Leskovec. Combining graph convolutional neural networks and label propagation. *ACM Transactions on Information Systems*, 40, 10 2022.
- [Wu17] Jianxin Wu. Introduction to convolutional neural networks. Technical report, National Key Lab for Novel Software Technology. Nanjing University, 2017.
- [WWD08] Sebastian Winter, Stefan Wagner, and Florian Deissenboeck. *A Comprehensive Model of Usability*, pages 106–122. 2008.
- [WYKN20] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53:1–34, 2020.
- [YB22] Emre Yalcin and Alper Bilge. Evaluating unfairness of popularity bias in recommender systems: A comprehensive user-centric analysis. *Information Processing and Management*, 59:103100, 11 2022.
- [YLS⁺11] Shuang-Hong Yang, Bo Long, Alexander J. Smola, Hongyuan Zha, and Zhaohui Zheng. Collaborative competitive filtering. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 295–304. ACM, 7 2011.
- [ZDN⁺08] Shiwan Zhao, Nan Du, Andreas Nauerz, Xiatian Zhang, Quan Yuan, and Rongyao Fu. Improved recommendation based on collaborative tagging behaviors. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 413–416. ACM, 1 2008.
- [ZYZ⁺22] Yihao Zhang, Meng Yuan, Chu Zhao, Mian Chen, and Xiaoyang Liu. Integrating label propagation with graph convolutional networks for recommendation. *Neural Computing and Applications*, 34:8211–8225, 5 2022.